UNIVERSITÉ DE GENÈVE
FACULTY OF SCIENCE
Department of Informatics

SIP stochastic information processing

Information Theory
for Data Science and Machine Learning

Slava Voloshynovskiy

Miscellanea

Lecturer:

Prof. **Slava Voloshynovskiy**, Stochastic Information Processing (SIP) Group, Department of Computer Science, UNIGE

E-mail: svolos@unige.ch; **Office: 215**

Teaching Assistants:

Guillaume Quetant, PhD student, Stochastic Information Processing (SIP) Group, Department of Computer Science, UNIGE

E-mail: Guillaume.Quetant@unige.ch

Mariia Drozdova, PhD student, Stochastic Information Processing (SIP) Group, Department of Computer Science, UNIGE

E-mail: mariia.drozdova@unige.ch

Organization and evaluation

Lectures

- Friday 10:00-12:00, Bat A/404-407, starting from February 25, 2022

HWs

Wednesday 8:00-10:00, Bat D/Amphi, starting from March 3, 2023

Web Site: moodle 12x004 Théorie de l'information pour la science des données et l'apprentissage automatique

Organization

- One semester
  - 14 lectures
  - 10 HWs

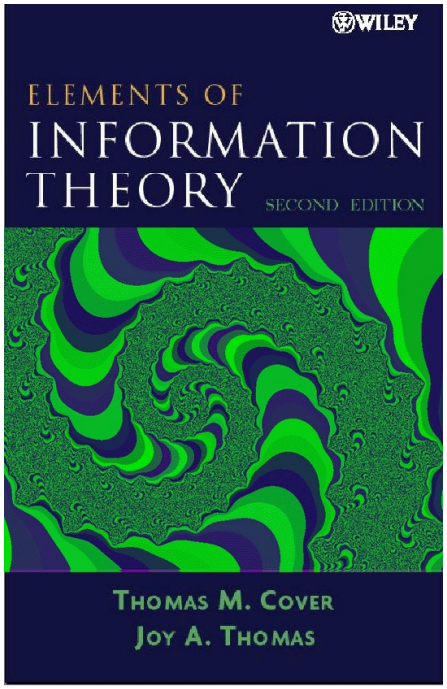
The final note is composed of

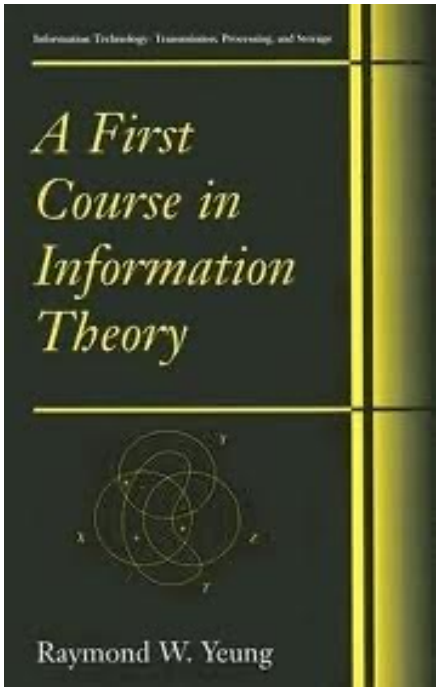
- Oral exam or CC (consisting of two parts during the semester) - 2/3
- HWs - 1/3

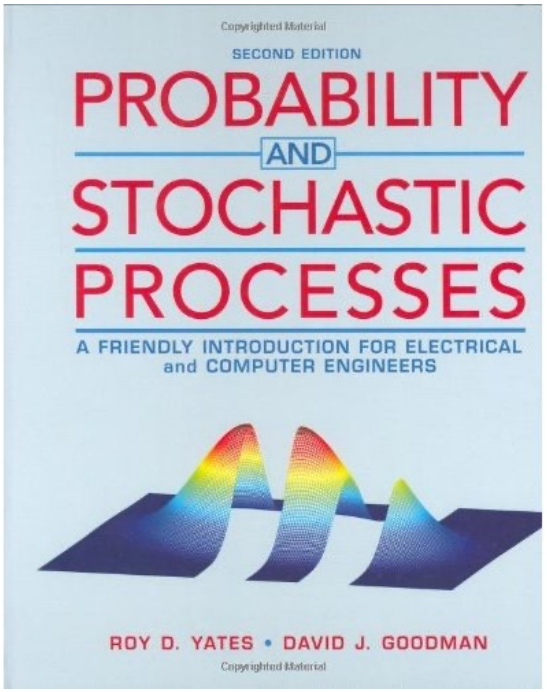
Copyright notice

This course uses in part some materials from previously mentioned recommended sources.

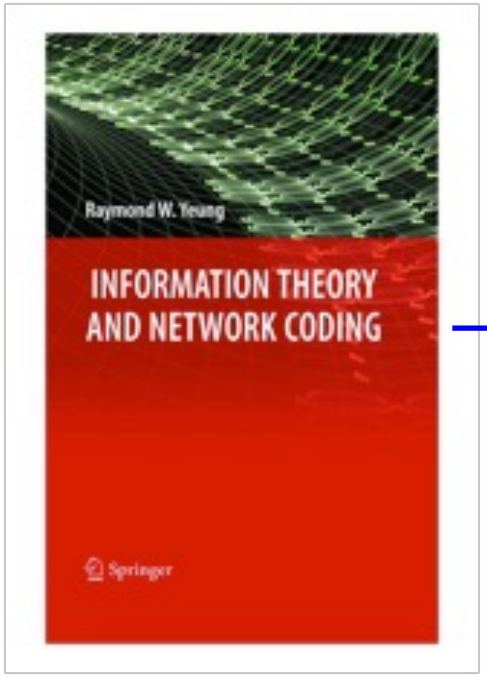
Recommended books



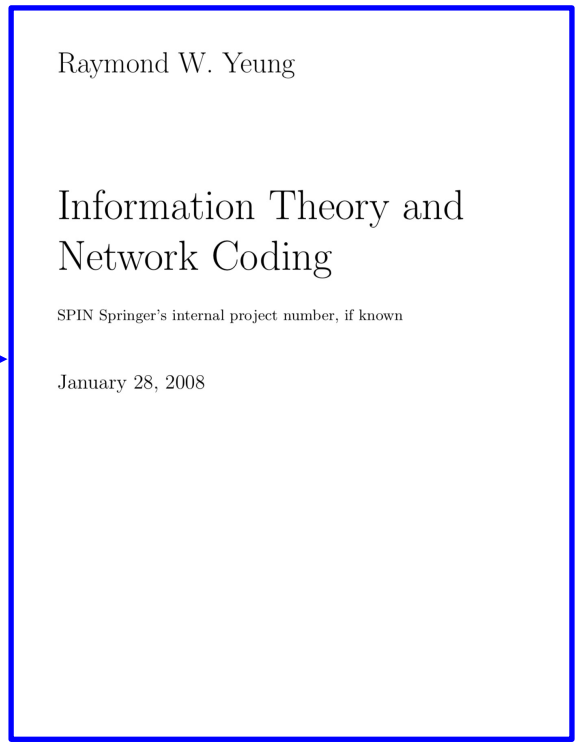




Recommended books



Free download as


Raymond W. Yeung

Information Theory and Network Coding

SPIN Springer's internal project number, if known

January 28, 2008

http://iest2.ie.cuhk.edu.hk/~whyeung/book2/

http://iest2.ie.cuhk.edu.hk/~whyeung/post/manuscript/main2.pdf

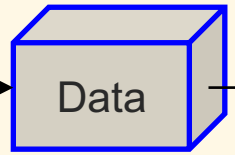
Content of course

- Theme 1 – Basic statistical data models
- Theme 2 – Information theoretic measures

Scope of class

IT measures

Statistical source of data

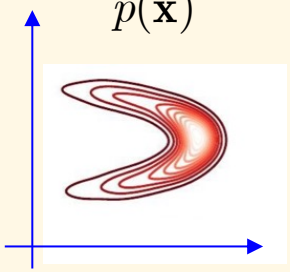


Data Science and Machine Learning

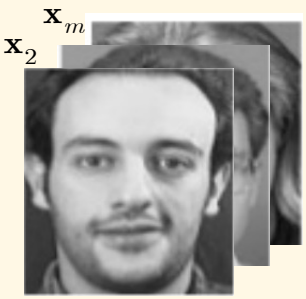
- Data analysis/visualization
- Data processing
  - clustering
  - denoising
  - restoration
  - classification
- Data compression
- Data transmission
- Data generation

Complex statistical distribution

$p(\mathbf{x})$



$\mathbf{X} =$



Content of this lecture

In this lecture we will:

- Introduce basic notations
- Recall basics of probability and matrix algebra
- Consider basic data models
- Recall transformation of random variables and vectors
- Recall weak law of large numbers and the central limit theorem (CLT)

Basic notions

- Random variables and vectors
  - Pmf/pdf
  - Moments
- Data representation
- Modeling of random vectors
  - Chain rule for probability
  - Independence and histograms
  - Markov chains
  - Conditional independence
  - Marginalization

Notations

- **Basic notations of scalars and vectors**

	Realization	Random
Scalar	x	X
Vector	$\mathbf{x} = x^N$	$\mathbf{X} = X^N$

$x_i \in \mathcal{X}$ or $x_i \in S_X$

- Discrete alphabets

$\mathcal{X} = \mathbb{Z}$ - integer

$\mathcal{X} = \{0,1\}$ - binary

$\mathcal{X} = \{a, b, \cdots, z\}$

$\mathcal{X} = \{January, February, \cdots, Decemeber\}$

- Continuous alphabets

$\mathcal{X} = \mathbb{R}$

- Vector $\mathbf{x} \triangleq \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} =$ 

$\mathbf{X} \triangleq \begin{bmatrix} | \\ \mathbf{X} \\ | \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}$

Notations

Matrix operations

$\mathbf{x}^T$	Transposed vector $\mathbf{x}$
$\mathbf{I}$	Identity matrix
$\mathbf{A}^T$	Transposed matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	Inverse matrix $\mathbf{A}$
$tr(\mathbf{A})$	Trace of matrix $\mathbf{A}$
$\det \lvert\mathbf{A}\rvert$	Determinant of matrix $\mathbf{A}$

Recall B: matrix operations

Random variables

- **Basic notations of probality**

$\Pr\left[X = x\right]$ Probability of event

  - Discrete random variables

$X \sim p_X(x)$ (or $X \sim p(x)$) Probability mass function (**pmf**) of discrete random variable X

$p_X(x \mid Y = y)$ (or $p_X(x \mid y)$) Conditional pmf of discrete random variable X given Y

  - Continuous random variables

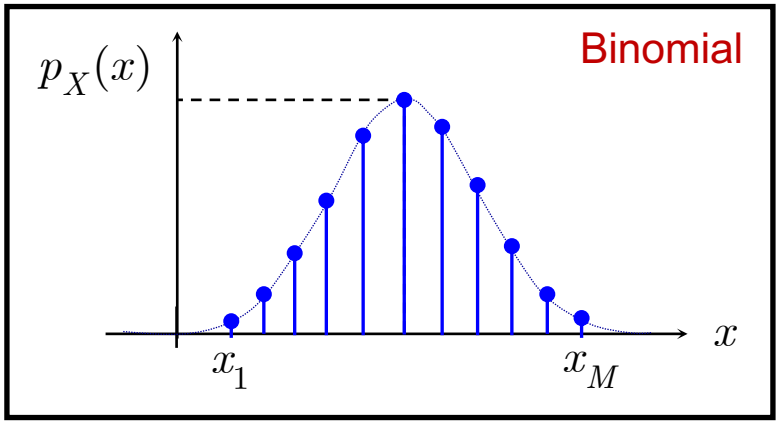
$X \sim f_X(x)$ (or $X \sim f(x)$) Probability density function (**pdf**) of continuous random variable X

$f_X(x \mid Y = y)$ (or $f_X(x \mid y)$) Conditional pdf of continuous random variable X given Y

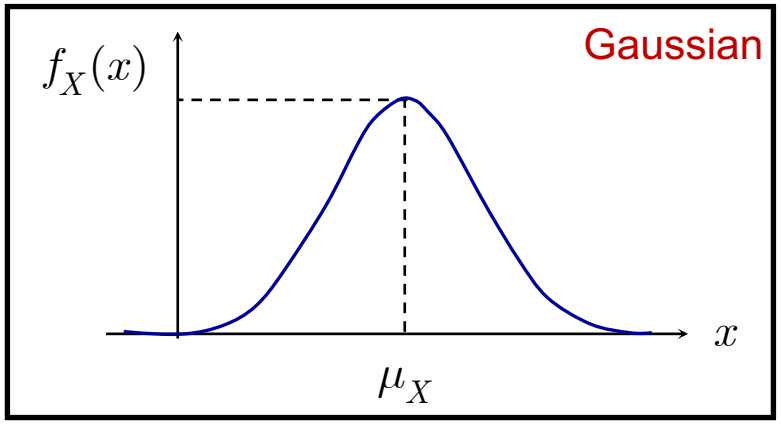
Random variables: pmf and pdf

- **Basic notations of probability**

  - Discrete random variables



  - Continuous random variables



$$p_X(x) = \Pr\left[X = k\right] = \begin{pmatrix} M \\ k \end{pmatrix} p^k \left(1-p\right)^{M-k} \qquad f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{\left(x-\mu_X\right)^2}{2\sigma_X^2}}$$

$$k = 0,1,\cdots,M$$

$$X \sim p_X(x) = \mathcal{B}(M,p) \qquad X \sim f_X(x) = \mathcal{N}(\mu_X, \sigma_X^2)$$

Random variables: moments

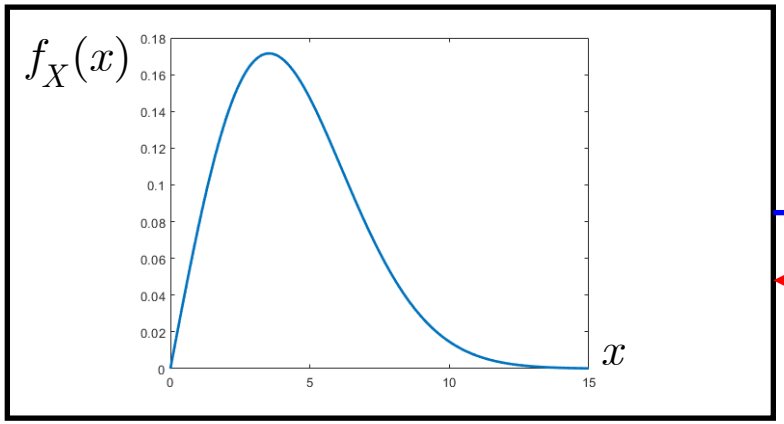
- **Moments**

	Discrete random variables	Continuous random variables
1st moment (**mean**)	$\mu_X = E_{p_X(x)}\left[X\right] = \sum_{i=1}^{M} x p(x_i)$	$\mu_X = E_{f_X(x)}\left[X\right] = \int_{x\in\mathcal{X}} x f_X(x)\,dx$
2nd moment (**variance**)	$\sigma_X^2 = Var\left[X\right] = E_{p_X(x)}\left[\left\|X-\mu_X\right\|^2\right] = \sum_{i=1}^{M}(x_i-\mu_X)^2 p(x_i)$	$\sigma_X^2 = Var\left[X\right] = E_{f_X(x)}\left[\left\|X-\mu_X\right\|^2\right] = \int_{x\in\mathcal{X}}(x-\mu_X)^2 f_X(x)\,dx$
3rd moment (**skewness**)	$E_{p_X(x)}\left[\left\|X-\mu_X\right\|^3\right]\Big/\sigma_X^3$	$E_{f_X(x)}\left[\left\|X-\mu_X\right\|^3\right]\Big/\sigma_X^3$
4th moment (**kurtosis**)	$E_{p_X(x)}\left[\left\|X-\mu_X\right\|^4\right]\Big/\sigma_X^4$	$E_{f_X(x)}\left[\left\|X-\mu_X\right\|^4\right]\Big/\sigma_X^4$

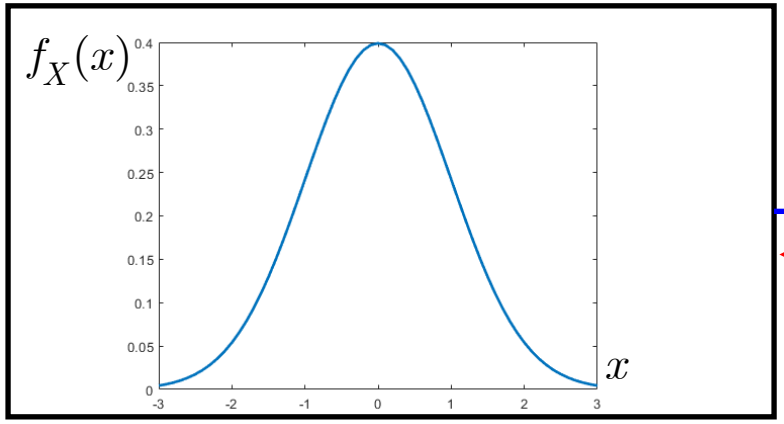
Random variables: moments

- General case: given a pmf/pdf, we can compute moments (not for all)

Any pdf



$$\mathcal{N}(\mu_X, \sigma_X^2)$$



$$\begin{bmatrix} E_{p_X(x)}\left[X\right] \\ E_{p_X(x)}\left[\left|X-\mu_X\right|^2\right] \\ E_{p_X(x)}\left[\left|X-\mu_X\right|^4\right] \\ E_{p_X(x)}\left[\left|X-\mu_X\right|^5\right] \\ \vdots \end{bmatrix}$$

$$\begin{bmatrix} E_{p_X(x)}\left[X\right] \\ E_{p_X(x)}\left[\left|X-\mu_X\right|^2\right] \end{bmatrix}$$

Intuitive interpretation

See Appendix D for more details

Basic operation: moments of sum of two variables

- Expected value of function of two random variables $Z = g(X,Y)$

$$E[Z] = \sum_{x \in S_X} \sum_{y \in S_Y} g(x,y) p_{X,Y}(x,y)$$

$$Z = g(X,Y) = X + Y$$

- **Mean** of sum of two random variables

$$E[Z] = E[X+Y] = E[X] + E[Y]$$

- **Variance** of sum of two random variables

$$Var[Z] = Var[X+Y] = Var[X] + Var[Y] + 2E[(X-\mu_X)(Y-\mu_Y)]$$

Remark: we will see pdf of sum of two and many variables later

Recall of variance and its properties

- Proof

$$Var\left[Z\right] = Var\left[X+Y\right] = E\left[\left(\left(X+Y\right)-\left(\mu_X+\mu_Y\right)\right)^2\right]$$

$$= E\left[\left(\left(X-\mu_X\right)+\left(Y-\mu_Y\right)\right)^2\right]$$

$$= E\left[\left(\left(X-\mu_X\right)^2+2\left(X-\mu_X\right)\left(Y-\mu_Y\right)+\left(Y-\mu_Y\right)^2\right)\right]$$

$$= E\left[\left(X-\mu_X\right)^2\right]+2E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]+E\left[\left(Y-\mu_Y\right)^2\right] =$$

$$= Var\left[X\right]+Var\left[Y\right]+2E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]$$

Random variables: covariance

- Covariance between two random variables

$$Cov\left[X,Y\right] = E_{p(x,y)}\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right] = \sum_{(x,y)\in S_{X,Y}} p(x,y)(x-\mu_X)(y-\mu_Y)$$

$$= E_{p(x,y)}\left[XY\right] - \mu_X\mu_Y = \sigma_{XY}$$

  - Variance of sum via covariance

$$Var\left[X+Y\right] = Var\left[X\right] + Var\left[Y\right] + 2Cov\left[X,Y\right]$$

  - Uncorrelated random variables

$$E\left[XY\right] = E\left[X\right]E\left[Y\right] = \mu_X\mu_Y$$

$$Cov\left[X,Y\right] = \mu_X\mu_Y - \mu_X\mu_Y = 0$$

$$\Rightarrow Var\left[X+Y\right] = Var\left[X\right] + Var\left[Y\right]$$

Random variables: covariance

- Correlation coefficient

$$\rho_{XY} = \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$$-1 \leq \rho_{XY} \leq 1$$

Random variables: summary

- Properties of variance and covariance
  - Let X and Y be random variables and a, b, c be constants.

$$Var\left[X\right] = E\left[\left|X - E\left[X\right]\right|^2\right] = E\left[X^2\right] - E\left[X\right]^2$$

$$0 \le Var\left[X\right] \le E\left[X^2\right]$$

$$Var\left[cX\right] = c^2 Var\left[X\right]$$

$$Var\left[aX + bY\right] = a^2 Var\left[X\right] + 2ab\,Cov\left[X,Y\right] + b^2 Var\left[Y\right]$$

$$Cov\left[X,Y\right] = E_{p(x,y)}\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right] = \sum_{(x,y)\in S_{X,Y}} p(x,y)(x - \mu_X)(y - \mu_Y)$$

$$= E_{p(x,y)}\left[XY\right] - \mu_X \mu_Y = \sigma_{XY}$$

See **Appendix A** for the recall, if needed

- From two variables to multidimensional vectors

Basic notions

- Random variables and vectors
  - Pmf/pdf
  - Moments
- Data representation
- Modeling of random vectors
  - Chain rule for probability
  - Independence and histograms
  - Markov chains
  - Conditional independence
  - Marginalization

Random vectors

- Vector $\mathbf{x} \triangleq \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$ random $\mathbf{X} \triangleq \begin{bmatrix} | \\ \mathbf{X} \\ | \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}$

$$\mathbf{X} \triangleq \begin{bmatrix} | \\ \mathbf{X} \\ | \end{bmatrix} = \begin{bmatrix} X_1 \\ X_i \\ \vdots \\ X_N \end{bmatrix} \begin{matrix} \rightarrow p_X(x_i) \\ \\ \\ \\ \rightarrow f_X(x_i) \end{matrix}$$

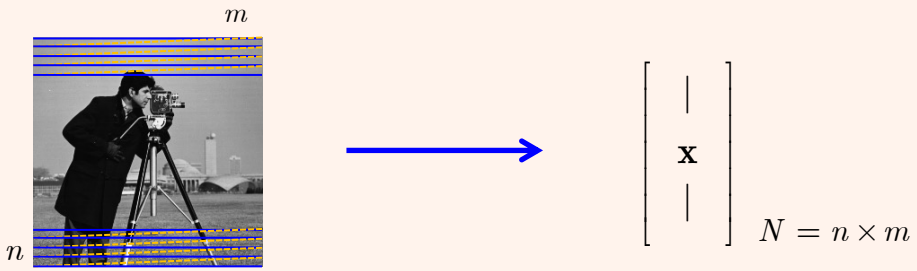
- Joint pmf or pdf

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{x}}(x_1, x_2, \cdots, x_N)$$

$$f_{\mathbf{x}}(\mathbf{x}) = f_{\mathbf{x}}(x_1, x_2, \cdots, x_N)$$

Data representation - 1

- **Example:**
  - Consider an image as a matrix
  - Read in a zig-zag order into one vector



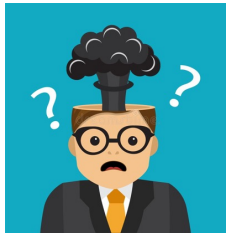
- Joint pmf defines all statistical relations between all elements of image/vector

$$\mathbf{X} = \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}$$

- Complexity
  - Modern image of size 1000x1000

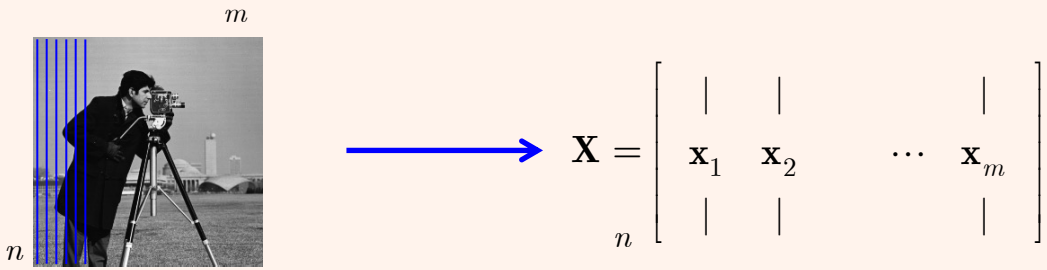
    $N = n \times m = 1000 \times 1000 = 1000000$

    All pair-wise, triple-wise, etc.



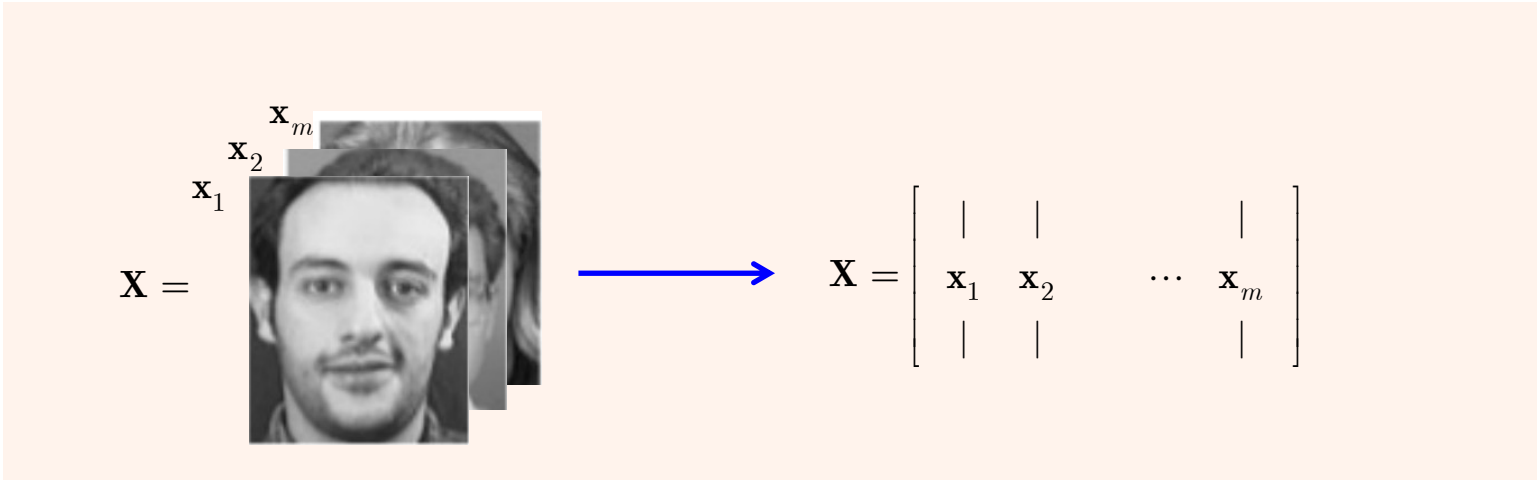
Data representation - 2

- **Example:**
  - Consider an image as a matrix
  - Read by rows as a matrix



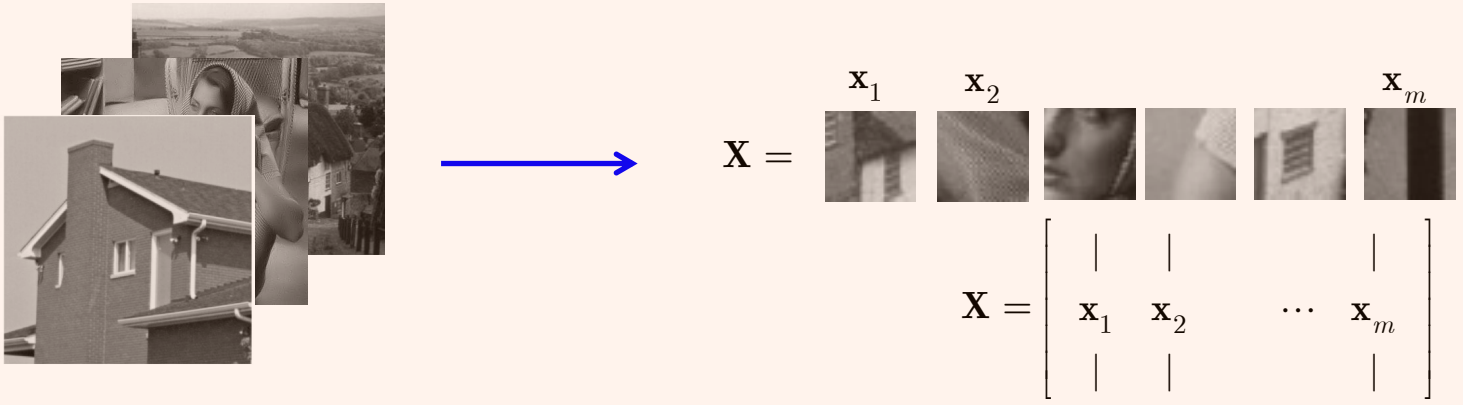
Data representation - 3

- **Example:**
  - image represents a column in a data matrix



Data representation - 4

- **Example**
  - image representation as patches



Basic notions

- Random variables and vectors
  - Pmf/pdf
  - Moments
- Data representation
- Modeling of random vectors
  - Chain rule for probability
  - Independence and histograms
  - Markov chains
  - Conditional independence
  - Marginalization

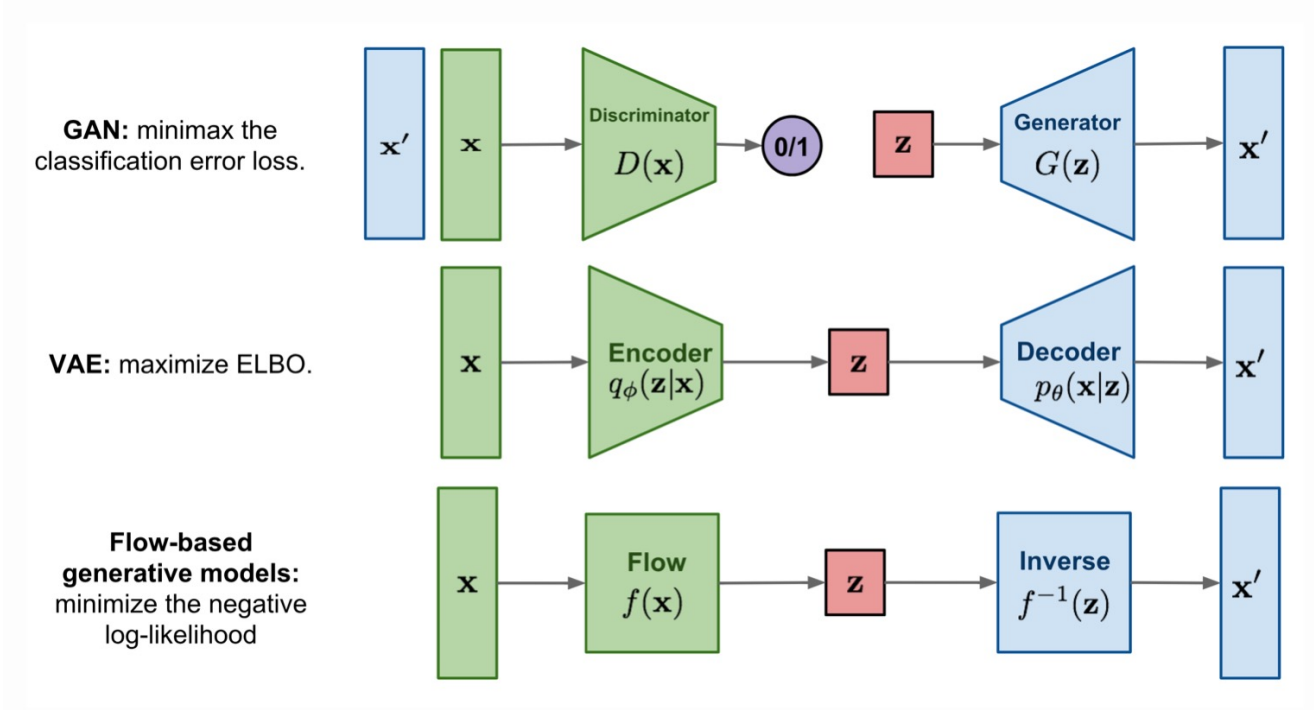
Random vectors: models

- Modeling of real data represented by random vectors is a very complex problem
  - Computational issues
  - Memory storage
  - Variability between data in various applications (medicine, astro, biometrics, etc.)

- Existing approaches to attack this issue are based on different assumptions:
  - Models based on independence
  - Models based on low-order dependencies /Markov chain/
  - Models based on second-order statistics /multivariate Gaussians/
  - Advanced models based on multi-level, multi-stage representations /deep networks/:
    - Variational autoencoder (VAE)
    - Generative adversarial networks (GAN)
    - Flow based models
    - Diffusion and score based models

Random vectors: sota models

- State of the art (sota) models



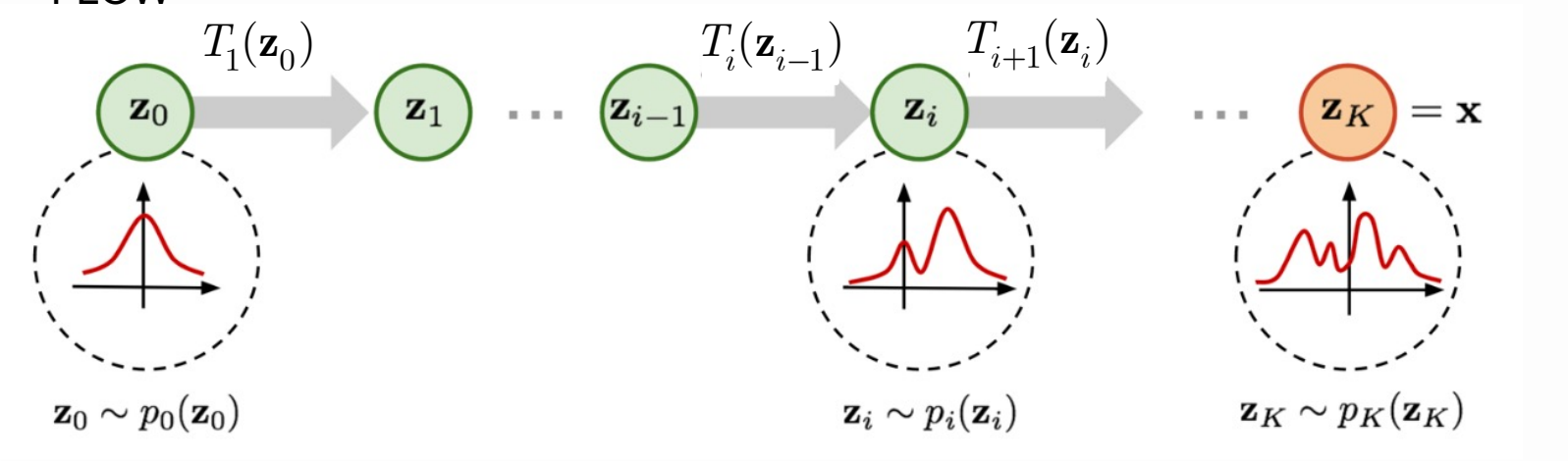
However, to understand these complex models we need first to recall simple ones.

https://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html

Random vectors: via transformation

- State of the art models

FLOW



Ilustration of a normalizing flow model, transforming a simple distribution to a complex one step by step

https://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html

Random vectors: chain rule decomposition

- Chain rule for probability: joint probability

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{x}}(x_1, x_2, \cdots, x_N) = p\left(x_1, x_2, ..., x_N\right) \quad \text{(Notations)}$$

$$p\left(x_1, x_2, ..., x_N\right) = p\left(x_1\right) p\left(x_2 \middle| x_1\right) p\left(x_3 \middle| x_2, x_1\right) ... p\left(x_N \middle| x_{N-1}, ..., x_2, x_1\right) =$$

$$= \prod_{i=1}^{N} p\left(x_i \middle| x_{i-1}, ..., x_2, x_1\right)$$

$$p\left(x_1, x_2\right) = p\left(x_1\right) p\left(x_2 \middle| x_1\right) = p\left(x_2\right) p\left(x_1 \middle| x_2\right)$$

$$\Rightarrow p\left(x_2 \middle| x_1\right) = \frac{p\left(x_2\right) p\left(x_1 \middle| x_2\right)}{p\left(x_1\right)}$$

Random vectors: independence

- Using the chain rule

$$p\left(x_1, x_2, \ldots, x_N\right) = p\left(x_1\right) p\left(x_2 \middle| x_1\right) p\left(x_3 \middle| x_2, x_1\right) \ldots p\left(x_N \middle| x_{N-1}, \ldots, x_2, x_1\right) = \prod_{i=1}^{N} p\left(x_i \middle| x_{i-1}, \ldots, x_2, x_1\right)$$

$$p\left(x_1, x_2\right) = p\left(x_1\right) p\left(x_2 \middle| x_1\right) = p\left(x_2\right) p\left(x_1 \middle| x_2\right) \underset{\substack{independent \\ X_1 \perp X_2}}{\Rightarrow} p\left(x_1\right) p\left(x_2\right)$$

$$p_{\mathbf{x}}\left(x_1, x_2, \ldots, x_N\right) = p_{x_1}\left(x_1\right) p_{x_2}\left(x_2\right) p_{x_3}\left(x_3\right) \ldots p_{x_N}\left(x_N\right) = \prod_{i=1}^{N} p_{x_i}\left(x_i\right)$$

For the independent and identically distributed (i.i.d.) random variables

$$p_{\mathbf{x}}\left(x_1, x_2, \ldots, x_N\right) = p_x\left(x_1\right) p_x\left(x_2\right) p_x\left(x_3\right) \ldots p_x\left(x_N\right) = \prod_{i=1}^{N} p_x\left(x_i\right)$$

Random vectors: independence

- Practical application of independence

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{x}}(x_1, x_2, \cdots, x_N)$$

$$\mathbf{X} = \quad \text{  }$$

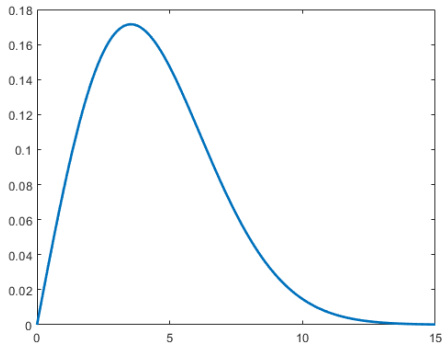
$$\mathbf{X} \triangleq \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}$$

Simplification

$$p\left(x_1, x_2, ..., x_N\right) = p\left(x_1\right) p\left(x_2\right) p\left(x_3\right) ... p\left(x_N\right)$$

$$\mathbf{X} \triangleq \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} \begin{matrix} \sim p\left(x_1\right) \\ \sim p\left(x_2\right) \\ \\ \sim p\left(x_N\right) \end{matrix}$$

$p(x_i)$



x_i

Matlab example of histogram of image pixels

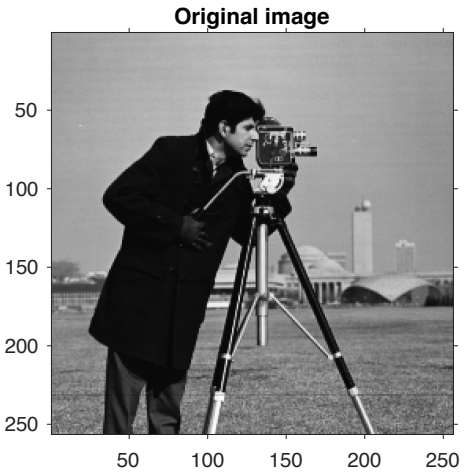
histo.m

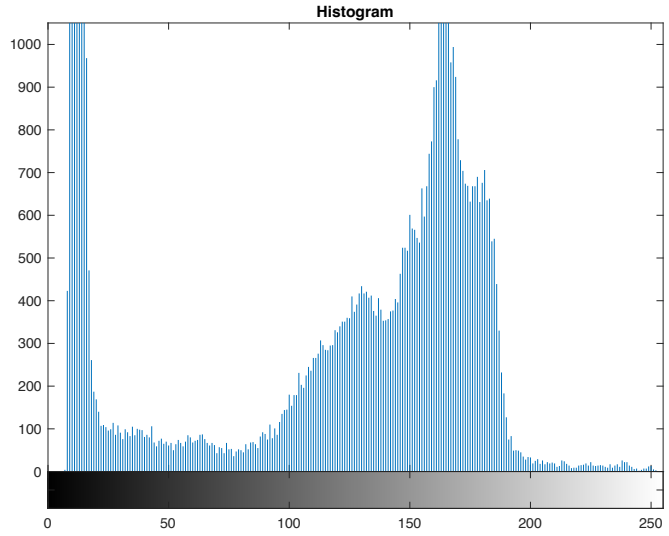
```
% Image histogram
% doc imhist

% Read image
im=imread('cameraman.tif');

% Show image
figure(1); imshow(im); title('Original image')

% Show histogram of image
figure(2); imhist(im); title('Histogram')
```

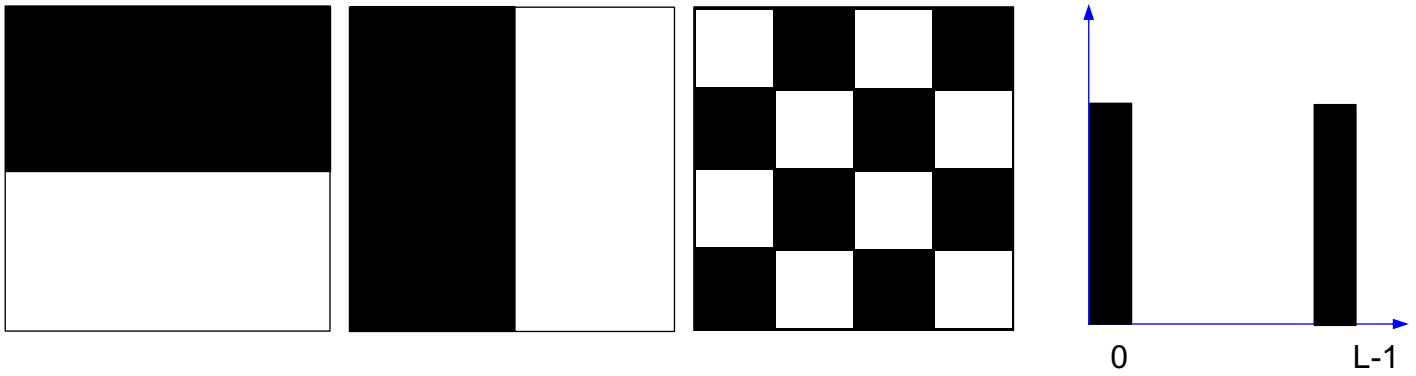




Remarks about histogram

Is histogram informative? Yes

Is histogram unique? Many images might have the same histogram!



Important: histogram is not a pdf or pmf!

Random vectors: moments

- If random variables X and Y are independent $\left(X \perp Y \right)$

$$E\left[XY \right] = E\left[X \right] E\left[Y \right]$$
$$E\left[X \middle| Y = y \right] = E\left[X \right]$$
$$E\left[Y \middle| X = x \right] = E\left[Y \right]$$
$$Cov\left[X, Y \right] = 0, \rho_{X,Y} = 0$$
$$Var\left[X + Y \right] = Var\left[X \right] + Var\left[Y \right]$$

- Proof:
  - For two independent random variables $p_{X,Y}\left(x, y \right) = p_X\left(x \right) p_Y\left(y \right)$

$$E\left[XY \right] = E\left[X \right] E\left[Y \right]$$

$$E[XY] = \sum_{x \in S_X} \sum_{y \in S_Y} xy p_X(x) p_Y(y) = \sum_{x \in S_X} x p_X(x) \sum_{y \in S_Y} y p_Y(y) = E[X] E[Y]$$

Indépendance entre v.a.

$$Cov\left[X,Y\right]=0,\rho_{X,Y}=0$$

$$Cov\left[X,Y\right]=E\left[XY\right]-\mu_X\mu_Y=E\left[X\right]E\left[Y\right]-\mu_X\mu_Y=0$$

$$\rho_{XY}=\frac{Cov\left[X,Y\right]}{\sqrt{Var\left[X\right]Var\left[Y\right]}}=0$$

$$Var\left[X+Y\right]=Var\left[X\right]+Var\left[Y\right]$$

$$Var\left[X\pm Y\right]=Var\left[X\right]+Var\left[Y\right]\pm 2\underbrace{Cov\left[X,Y\right]}_{=0}$$

Random vectors: Markovianity

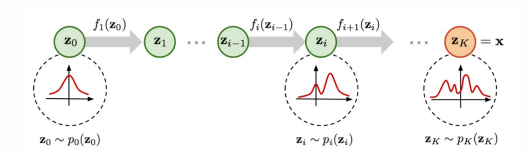
- The **first order** Markov model
  - assumes the dependence on one element from past

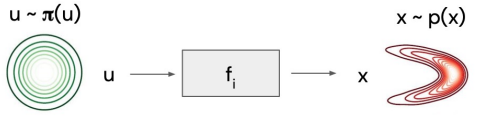
$$p\left(x_i \middle| x_{i-1}, x_{i-2}, \cdots x_1\right) \underset{1st \text{ order Markov model}}{\Rightarrow} p\left(x_i \middle| x_{i-1}\right)$$

$$p\left(x_1, x_2, \ldots, x_N\right) = p\left(x_1\right) p\left(x_2 \middle| x_1\right) p\left(x_3 \middle| x_2, x_1\right) \ldots p\left(x_N \middle| x_{N-1}, \ldots, x_2, x_1\right) =$$
$$\underset{1st \text{ order Markov model}}{\Rightarrow} p\left(x_1\right) p\left(x_2 \middle| x_1\right) p\left(x_3 \middle| x_2\right) \ldots p\left(x_N \middle| x_{N-1}\right)$$

$$p\left(x_1, x_2, \ldots, x_N\right) \underset{1st \text{ order Markov model}}{\Rightarrow} p\left(x_1\right) \prod_{i=2}^{N} p\left(x_i \middle| x_{i-1}\right)$$

The modeling of a complex pmf/pdf is reduced to the definition of pair-wise conditional distributions





HAMILTONIAN GENERATIVE NETWORKS

Random vectors: Markovianity

- The **second order** Markov model
  - assumes the dependence on two elements from past

$$p\left(x_i \middle| x_{i-1}, x_{i-2}, \cdots x_1\right) \underset{2nd \text{ order Markov model}}{\Rightarrow} p\left(x_i \middle| x_{i-1}, x_{i-2}\right)$$

$$p\left(x_1, x_2, \ldots, x_N\right) = p\left(x_1\right) p\left(x_2 \middle| x_1\right) p\left(x_3 \middle| x_2, x_1\right) \ldots p\left(x_N \middle| x_{N-1}, \ldots, x_2, x_1\right) =$$
$$\underset{2nd \text{ order Markov model}}{\Rightarrow} p\left(x_1\right) p\left(x_2 \middle| x_1\right) p\left(x_3 \middle| x_2, x_1\right) \ldots p\left(x_N \middle| x_{N-1}, x_{N-2}\right)$$

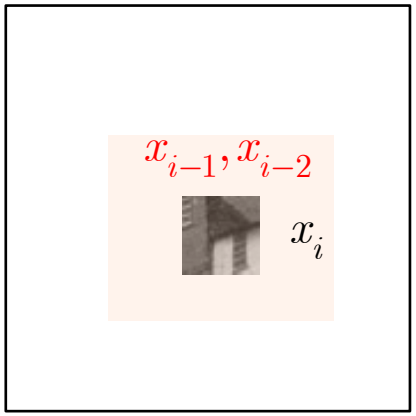
$$p\left(x_1, x_2, \ldots, x_N\right) \underset{2nd \text{ order Markov model}}{\Rightarrow} p\left(x_1\right) \prod_{i=2}^{N} p\left(x_i \middle| x_{i-1}, x_{i-2}\right)$$

Random vectors: Markovianity

- Practical application application of Markovianity for image processing

$$\mathbf{X} = \quad \mathbf{X} \triangleq \begin{bmatrix} | \\ \mathbf{X} \\ | \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}$$

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{x}}(x_1, x_2, \cdots, x_N)$$



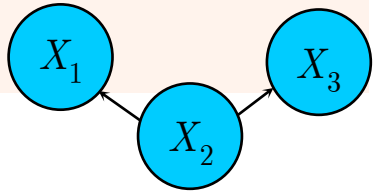
$$\mathbf{X} \triangleq \begin{bmatrix} | \\ \mathbf{X} \\ | \end{bmatrix} = \begin{bmatrix} \vdots \\ X_{i-2} \\ X_{i-1} \\ X_i \\ \vdots \end{bmatrix}$$

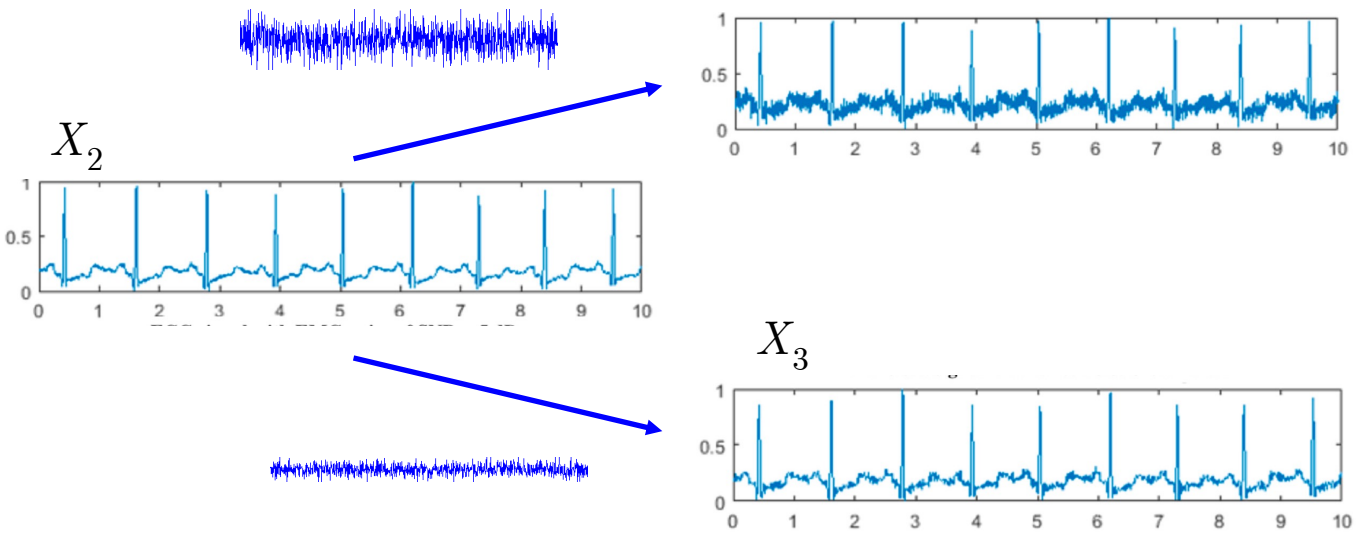
$$\prod_{i=2}^{N} p\left(x_i \middle| x_{i-1}, x_{i-2}\right)$$

Random vectors: conditional independence

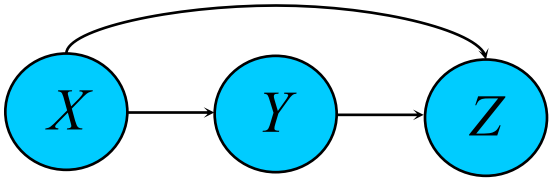
- Random variables X_1 and X_3 are independent given X_2

$$p(x_1, x_2, x_3) = p(x_2)\underbrace{p(x_1, x_3 | x_2)}_{=p(x_1|x_2)p(x_3|x_1,x_2)} = p(x_2)p(x_1|x_2)p(x_3|x_2)$$

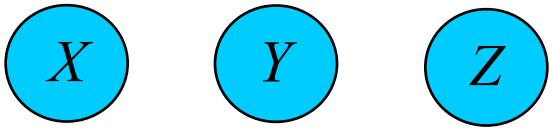




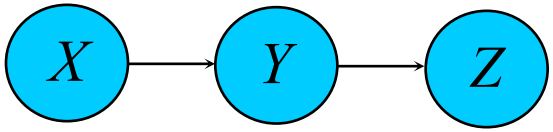
Summary



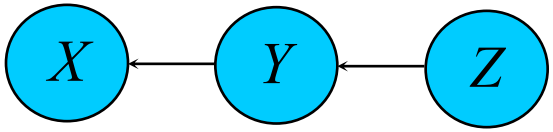
$$p(x,y,z) = p(x)\,p(y|x)\,p(z|x,y)$$



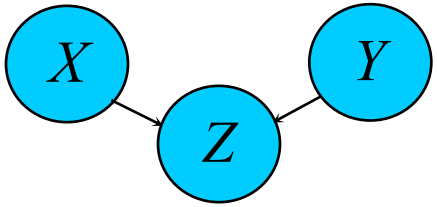
$$p(x,y,z) = p(x)\,p(y)\,p(z)$$



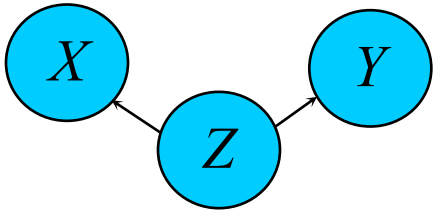
$$p(x,y,z) = p(x)\,p(y|x)\,p(z|y)$$



$$p(x,y,z) = p(z)\,p(y|z)\,p(x|y)$$



$$p(x,y,z) = p(z|x,y)\,p(x)\,p(y)$$



$$p(x,y,z) = p(z)\,p(x|z)\,p(y|z)$$

Random vectors: marginalization

- Marginalization:
  - Given $p\left(x_1, x_2\right)$, one can find marginal distributions by

$$p\left(x_1\right) = \sum_{x_2 \in S_{X_2}} p\left(x_1, x_2\right) = \sum_{x_2 \in S_{X_2}} p\left(x_1\right) p\left(x_2 \middle| x_1\right)$$

$$p\left(x_2\right) = \sum_{x_1 \in S_{X_1}} p\left(x_1, x_2\right) = \sum_{x_1 \in S_{X_1}} p\left(x_2\right) p\left(x_1 \middle| x_2\right)$$

Transformation of random variables

- General linear transform:
  - Random scalar variables
  - Random vectors
- Sum of random variables
  - Sum of independent random variables
  - Concentration inequalities
  - Weak Law of Large Numbers and CLT
  - Application to vector norms (see later after norms)
- Transformation of Gaussian random vectors
  - Mean and covariance matrix after transformation
  - Pdf after transformation
  - Properties of transform
    - Decorrelation
    - Whitening

Transformation of random variables

- Why are we interested in transformation of random variables/vectors?

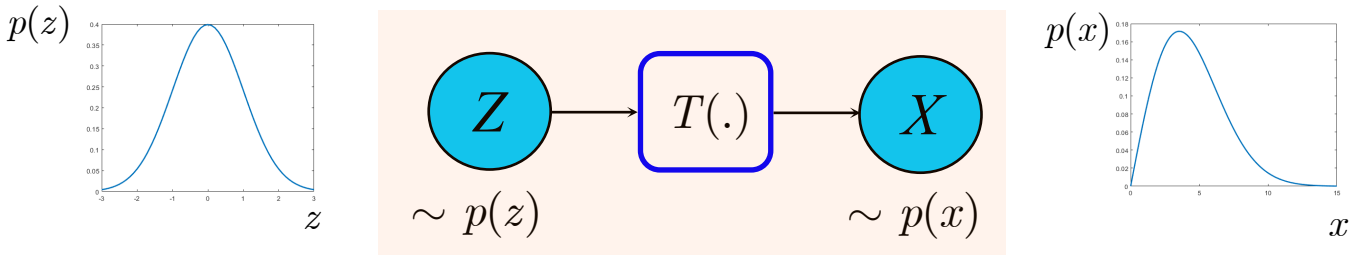
- The transformation is a core operation for:
  - Data analysis and visualization (looking for interesting features in data)
  - Fast data indexing and searching (lower dim data)
  - Data processing
    - Clustering (grouping of similar data), denoising (removal of noise), restoration (recovery from distortion), classification (assigning to the predefined class labels), etc.
  - Data compression (more efficient/compact data representation)
  - Data transmission (robust data encoding)
  - Data generation (generation of complex distributions from simple ones)

Transformation of random variables

- **Goal**: to transform a random variable $Z \sim p(z)$ into another random variable $X \sim p(x)$
  - It is assumed that Z follows some simple distribution that can be easily generated and analyzed (uniform or Gaussian)
  - While the distribution of X is complex (asymmetric, multimodal, etc.).
  - SOTA generative models (VAE, GAN, FLOW)

- **Scalar-to-scalar:** $T : \mathbb{R} \rightarrow \mathbb{R}$

One uses a transform $T(.)$



Note: formally we need to use $f_Z(z)$ instead of $p(z)$ for continuous variables

Transformation of random variables

- Properties of transform $T(.)$:
  - Invertible T^{-1}
  - Both T and T^{-1} are differentiable

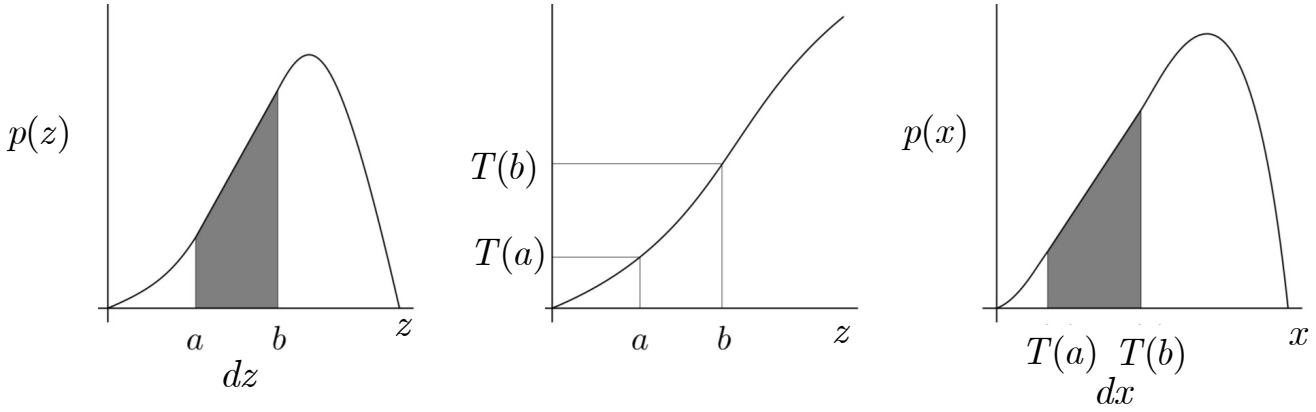
- Generate $Z \sim p(z)$
- Consider one-to-one mapping transform: $x = T(z)$
  - Recalling $\int p(z)dz = \int p(x)dx = 1$
  - Preservation of "volume":
    - Probability is invariant under the change of variables $\left|p(x)dx\right| = \left|p(z)dz\right|$

$$p(x) = p(z)\left|\frac{dz}{dx}\right| = p(T^{-1}(x))\left|\frac{dT^{-1}(x)}{dx}\right|$$

$$x = T(z)$$
$$z = T^{-1}(x)$$

Transformation of random variables

- Example (more details)



$$\Pr\big[a \le Z < b\big] = \int_a^b p(z)\,dz = \Pr\big[T(a) \le X < T(b)\big] = \int_{T(a)}^{T(b)} p(x)\,dx$$

$$\big|p(z)dz\big| = \big|p(x)dx\big|$$

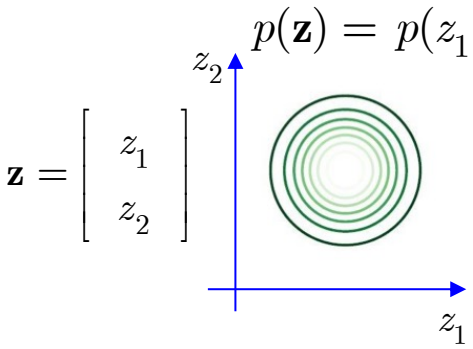
$$p(x) = p(z)\left|\frac{dz}{dx}\right|$$

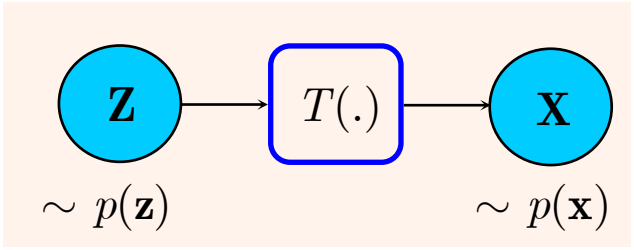
Transformation of random variables

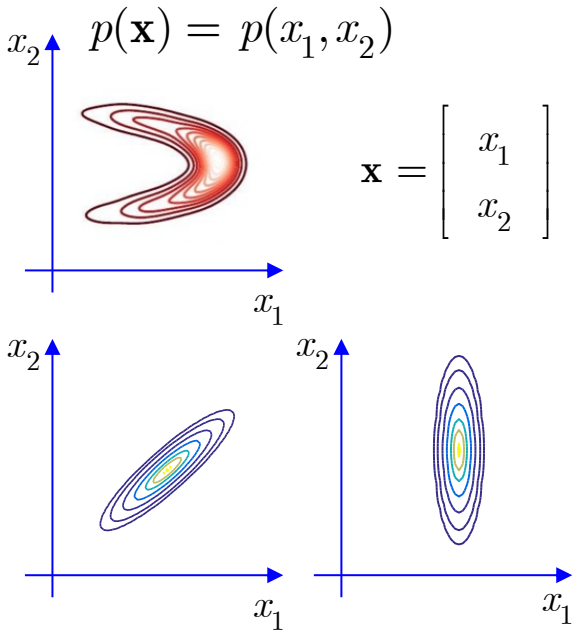
- **Vector-to-vector:** $T : \mathbb{R}^N \rightarrow \mathbb{R}^N$

$$\mathbf{x} = T(\mathbf{z})$$
$$\mathbf{z} = T^{-1}(\mathbf{x})$$

$$\begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} = T\left(\begin{bmatrix} | \\ \mathbf{z} \\ | \end{bmatrix}\right) = \begin{bmatrix} T_1(\mathbf{z}) \\ T_2(\mathbf{z}) \\ \vdots \\ T_N(\mathbf{z}) \end{bmatrix}$$







Transformation of random vectors

- **Vector-to-vector:** $T : \mathbb{R}^N \rightarrow \mathbb{R}^N$

$$\mathbf{x} = T(\mathbf{z})$$

$\mathbf{z} = T^{-1}(\mathbf{x})$, i.e., the transform is invertible

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{z}}(\mathbf{z})\left|\det J_T(\mathbf{z})\right|^{-1} \quad \Rightarrow \quad p_{\mathbf{x}}(\mathbf{x}) = \frac{p_{\mathbf{z}}\left(T^{-1}(\mathbf{x})\right)}{\left|\det J_T(\mathbf{z})\right|}$$

The Jacobian matrix:

$$J_T(\mathbf{z}) = \begin{bmatrix} \frac{\partial T_1(\mathbf{z})}{\partial z_1} & \cdots & \frac{\partial T_1(\mathbf{z})}{\partial z_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial T_N(\mathbf{z})}{\partial z_1} & \cdots & \frac{\partial T_N(\mathbf{z})}{\partial z_N} \end{bmatrix} \quad J_T(\mathbf{x}) = \begin{bmatrix} \frac{\partial T_1^{-1}(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial T_1^{-1}(\mathbf{x})}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial T_N^{-1}(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial T_N^{-1}(\mathbf{x})}{\partial x_N} \end{bmatrix}$$

Transformation of random variables

- Example: $T : \mathbb{R}^2 \to \mathbb{R}^2$

$$\mathbf{x} = \begin{bmatrix} T_1(\mathbf{z}) \\ T_2(\mathbf{z}) \end{bmatrix} = \begin{bmatrix} T_1(z_1, z_2) \\ T_2(z_1, z_2) \end{bmatrix} = \begin{bmatrix} z_1^2 z_2 \\ 5z_1 + \sin(z_2) \end{bmatrix}$$

The Jacobian matrix:

$$J_T(\mathbf{z}) = \begin{bmatrix} \dfrac{\partial T_1(\mathbf{z})}{\partial z_1} & \dfrac{\partial T_1(\mathbf{z})}{\partial z_2} \\ \dfrac{\partial T_2(\mathbf{z})}{\partial z_1} & \dfrac{\partial T_2(\mathbf{z})}{\partial z_2} \end{bmatrix} = \begin{bmatrix} 2z_1 z_2 & z_1^2 \\ 5 & \cos(z_2) \end{bmatrix}$$

The Jacobian determinant:

$$\left|\det J_T(\mathbf{z})\right| = 2z_1 z_2 \cos(z_2) - 5z_1^2$$

Linear transformation of random vectors

- **Vector-to-vector:** $T : \mathbb{R}^N \rightarrow \mathbb{R}^N$

$$\mathbf{x} = T(\mathbf{z}) = \mathbf{A}\mathbf{z}$$

$\mathbf{z} = T^{-1}(\mathbf{x}) = \mathbf{A}^{-1}\mathbf{x}$, i.e., the transform is invertible

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{p_{\mathbf{z}}\left(T^{-1}(\mathbf{x})\right)}{\left|\det J_T(\mathbf{z})\right|} = \frac{p_{\mathbf{z}}\left(\mathbf{A}^{-1}\mathbf{x}\right)}{\left|\det J_T(\mathbf{z})\right|}$$

The Jacobian matrix:

$$J_T(\mathbf{z}) = \begin{bmatrix} \frac{\partial T_1(\mathbf{z})}{\partial z_1} & \cdots & \frac{\partial T_1(\mathbf{z})}{\partial z_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial T_N(\mathbf{z})}{\partial z_1} & \cdots & \frac{\partial T_N(\mathbf{z})}{\partial z_N} \end{bmatrix}$$

Linear transformation of random vectors

- **Vector-to-vector:** $T : \mathbb{R}^N \to \mathbb{R}^N$

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{p_{\mathbf{z}}\left(T^{-1}(\mathbf{x})\right)}{\left|\det J_T(\mathbf{z})\right|} = \frac{p_{\mathbf{z}}\left(\mathbf{A}^{-1}\mathbf{x}\right)}{\left|\det J_T(\mathbf{z})\right|}$$

The Jacobian matrix:

$$\mathbf{x} = \mathbf{A}\mathbf{z} = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_N \end{bmatrix}$$

$$x_j = T_j(\mathbf{z}) = a_{j1}z_1 + \cdots + a_{ji}z_i + \cdots + a_{jN}z_N \qquad \frac{\partial T_j(\mathbf{z})}{\partial z_i} = \frac{\partial x_j}{\partial z_i} = a_{ji} \qquad \Rightarrow J_T(\mathbf{z}) = \mathbf{A}$$

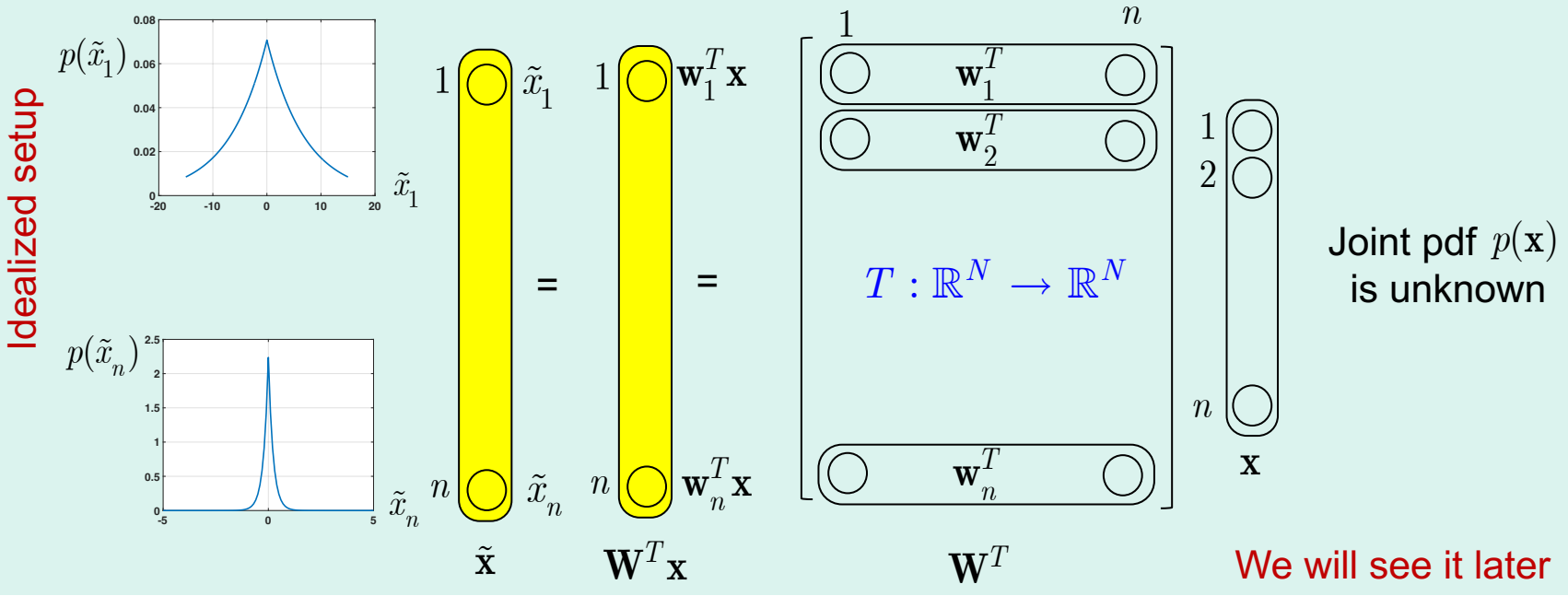
$$p_{\mathbf{x}}(\mathbf{x}) = \frac{p_{\mathbf{z}}\left(T^{-1}(\mathbf{x})\right)}{\left|\det J_T(\mathbf{z})\right|} = \frac{p_{\mathbf{z}}\left(\mathbf{A}^{-1}\mathbf{x}\right)}{\left|\det \mathbf{A}\right|}$$

Transformation for decorrelation and compaction

- **Main idea**: assume $\mathbf{x}$ can be transformed to some domain where all pixels will be independent (ideally; just a wish), i.e.:

$$p(\mathbf{x}) = p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2|x_1)\cdots p(x_n|x_{n-1}, \cdots, x_1) \Rightarrow p(\tilde{\mathbf{x}}) = p(\tilde{x}_1)p(\tilde{x}_2)\cdots p(\tilde{x}_n)$$

- such that we can uniquely transform back to the original vector



Transformation for decorrelation and compaction

- The key element of transformation is a projection

$$\mathbf{w}_i^T \mathbf{x} = \sum_{j=1}^{N} w_{i,j} x_j = w_{i,1} x_1 + w_{i,2} x_2 + \cdots w_{i,N} x_N$$

- Assuming $\left\{ w_{i,j} \right\}_{j=1}^{N} = 1$, we have just a sum

$$s_N(x) = x_1 + x_2 + \cdots + x_N$$

- Consider all variables to be random ones, we obtain a random value of sum

$$S_N\left(X\right) = X_1 + X_2 + \cdots + X_N$$

- Since the sum of random variables plays a fundamental role for many concentration properties, we will consider it next.

Transformation of random variables

- General linear transform:
  - Random scalar variables
  - Random vectors
- Sum of random variables
  - Sum of independent random variables
  - Concentration inequalities
  - Weak Law of Large Numbers and CLT
  - Application to vector norms (see later after norms)
- Transformation of Gaussian random vectors
  - Mean and covariance matrix after transformation
  - Pdf after transformation
  - Properties of transform
    - Decorrelation
    - Whitening

Sum of random variables: pdf and moments

- Sum of random variables $S_N(X) = X_1 + X_2 + \cdots + X_N$
  - It is also often of interest to compute a sampling mean

$$M_N(X) = \frac{S_N(X)}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

- The statistical behavior of $S_N(X)$ is characterized by:
  - **Pdf** $f_{X_1+\ldots+X_N}(x)$
  - If it is difficult to compute pdf, one computes the first two moments
    - **Expected value of sum**

$$E[S_N(X)] = E[X_1 + \ldots + X_N] = E[X_1] + \ldots + E[X_N]$$

    - **Variance of sum**

$$Var[S_N(X)] = \sum_{i=1}^{N} Var[X_i] + \sum_{i=1}^{N}\sum_{j\neq i}^{N} Cov[X_i, X_j]$$

Sum of random variables: pdf and moments

- Sum of uncorrelated random variables
  - Variance

$$Cov\left[X_i, X_j\right] = 0$$

$$Var\left[S_N(X)\right] = \sum_{i=1}^{N} Var\left[X_i\right] + \sum_{i=1}^{N}\sum_{j=1}^{N} \underbrace{Cov\left[X_i, X_j\right]}_{=0} = \sum_{i=1}^{N} Var\left[X_i\right]$$

- Sum of independent random variables

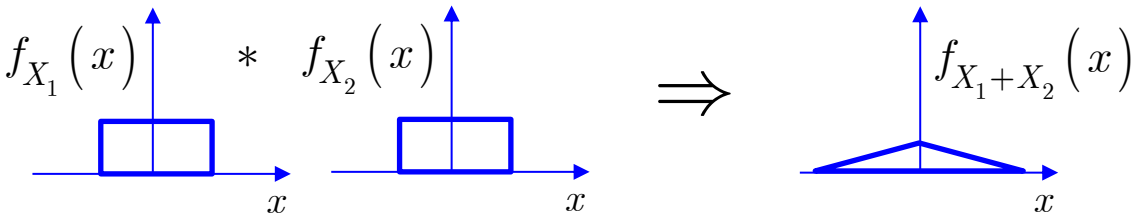
$$f_{X_1+\ldots+X_N}(x) = \left(f_{X_1}(x) * \ldots * f_{X_N}(x)\right)(x)$$

where $$f_{X_1+X_2}(x) = \int_{-\infty}^{+\infty} f_{X_1}(u) f_{X_2}(x-u) du = \left(f_{X_1}(x) * f_{X_2}(x)\right)(x)$$

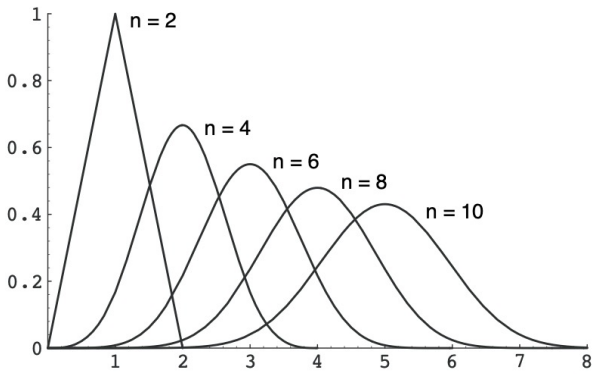
convolution

Sum of random variables: convergence to Gaussian

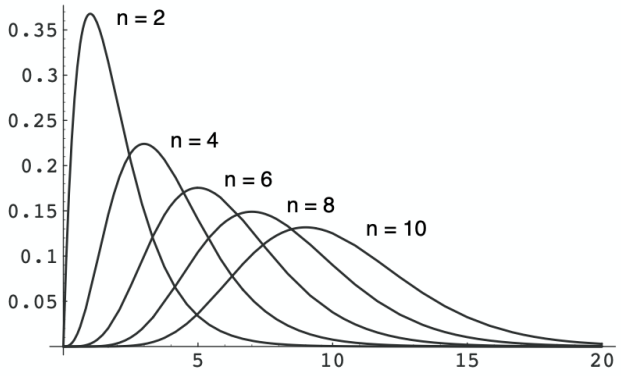
- Sum of independent random variables
  - Example: sum of two uniform independent random variables



Convolution of $N = n$ uniform pdfs



Convolution of $N = n$ exponential pdfs



- The pdf of sum converges to Gaussian pdf

Sample mean

- Consider a sample mean of N independent and identically distributed random variables $X_1, X_2, \cdots, X_N$ with $E\left[X_i\right] = \mu_X$ and $Var\left[X_i\right] = \sigma_X^2$

- The **sample mean** is

$$M_N\left(X\right) = \frac{S_N\left(X\right)}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

- The moments of **sample mean**
  - **Expected value:** $E\left[M_N\left(X\right)\right] = E\left[X\right] = \mu_X$
  - **Variance:** $Var\left[M_N\left(X\right)\right] = \frac{Var\left[X\right]}{N} = \frac{\sigma_X^2}{N}$

Sample mean: moments

- **Proof**

- **Expected value**: sampling mean is an unbiased estimator

$$E\left[M_N(X)\right] = \frac{1}{N}\left(E\left[X_1\right] + \cdots + E\left[X_N\right]\right) = \frac{1}{N}\left(E\left[X\right] + \cdots + E\left[X\right]\right) = E\left[X\right] = \mu_X$$

- **Variance**: suppose that $Var\left[X_i\right] = \sigma_X^2$

$$Var\left[M_N(X)\right] = \frac{1}{N^2}\left(Var\left[X_1\right] + \cdots + Var\left[X_N\right]\right)$$

$$= \frac{1}{N^2}\left(Var\left[X\right] + \cdots + Var\left[X\right]\right) = \frac{1}{N}Var\left[X\right] = \frac{\sigma_X^2}{N}$$

The variance of sample mean approaches zero as the number of samples increases (a.k.a. efficient estimator)

Weak Law of Large Numbers (WLLN)

- Let $X_1, X_2, \cdots, X_N$ be a sequence of i.i.d. random variables with finite mean $E\left[X_i\right] = \mu_X$ and variance $Var\left[X_i\right] = \sigma_X^2$

- For any $\varepsilon > 0$, **Chebyshev inequality** states:

$$(a)\ \Pr\left[\left|M_N\left(X\right) - \mu_X\right| \geq \varepsilon\right] \leq \frac{\sigma_X^2}{N\varepsilon^2}$$

$$(b)\ \Pr\left[\left|M_N\left(X\right) - \mu_X\right| < \varepsilon\right] \geq 1 - \frac{\sigma_X^2}{N\varepsilon^2}$$

- **Weak Law of Large Numbers**

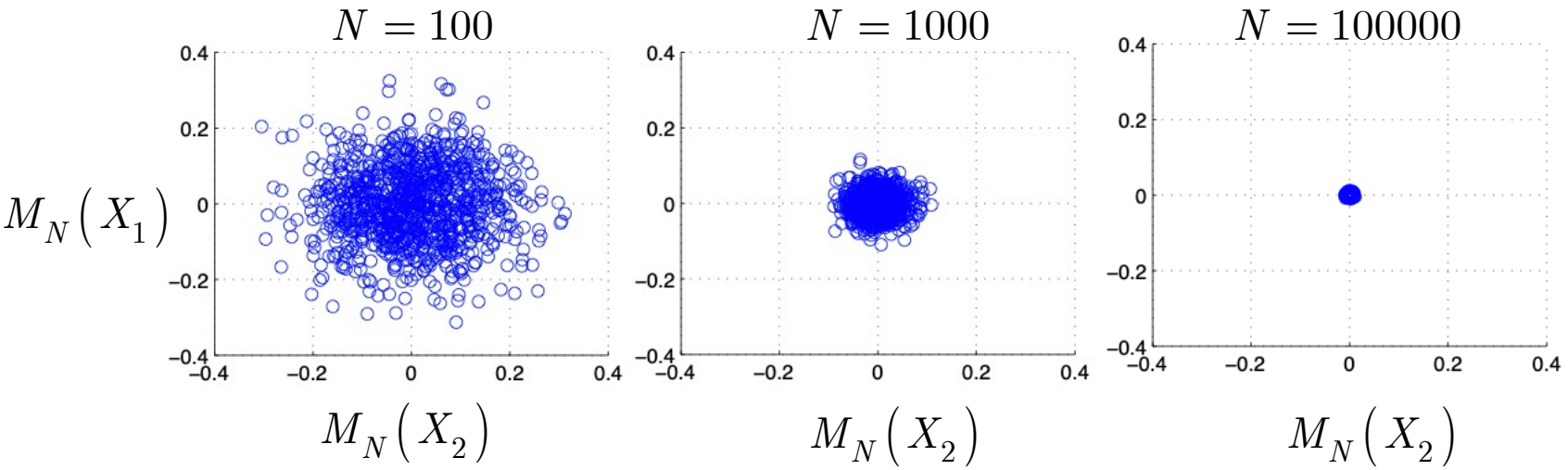
$$(a)\ \lim_{N\to\infty} \Pr\left[\left|M_N\left(X\right) - \mu_X\right| \geq \varepsilon\right] = 0$$

$$(b)\ \lim_{N\to\infty} \Pr\left[\left|M_N\left(X\right) - \mu_X\right| < \varepsilon\right] = 1$$

Weak Law of Large Numbers (WLLN)

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \qquad X_1 \perp X_2 \sim f_X(x) = \mathcal{N}(0, \sigma_X^2)$$

$$M_N(X_1) = \frac{S_N(X_1)}{N} = \frac{1}{N}\sum_{i=1}^{N} X_{1,i} \underset{N\to\infty}{\to} 0 \qquad M_N(X_2) = \frac{S_N(X_2)}{N} = \frac{1}{N}\sum_{i=1}^{N} X_{2,i} \underset{N\to\infty}{\to} 0$$

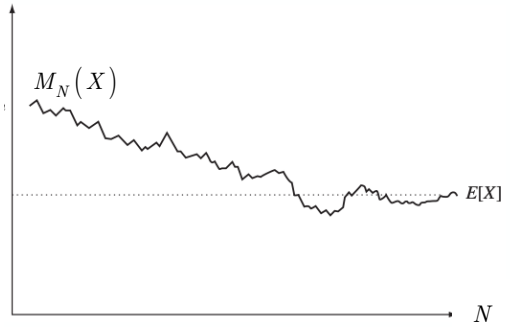


Strong Law of Large Numbers (SLLN)

- Let $X_1, X_2, \cdots, X_N$ be a sequence of i.i.d. random variables with finite mean $E\left[X_i\right] = \mu_X$ and variance $Var\left[X_i\right] = \sigma_X^2$ (some bounded)

- In this case

$$\Pr\left[\lim_{N\to\infty} M_N\left(X\right) = \mu_X\right] = 1$$



- $M_K\left(X\right)$ is the sample mean computed using K variables

- With probability 1, every sample mean calculation will eventually approach and stay close to $E\left[X_i\right] = \mu_X$

- The strong law implies the weak law

Central Limit Theorem (CLT)

- Let $X_1, X_2, \cdots, X_N$ be a sequence of i.i.d. random variables with finite mean $E\left[X_i\right] = \mu_X$ and variance $Var\left[X_i\right] = \sigma_X^2$

- Let $S_N\left(X\right)$ be the sum:

$$S_N\left(X\right) = X_1 + X_2 + \cdots + X_N$$

- Define the centered and normalized variable

$$Z_N\left(X\right) = \frac{S_N\left(X\right) - N\mu_X}{\sigma_X\sqrt{N}}$$

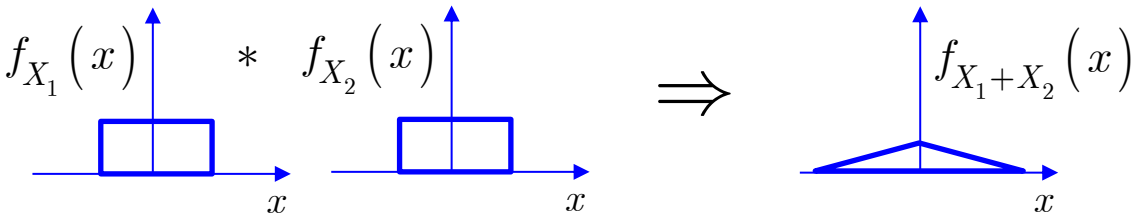
- then

$$\lim_{N\to\infty} \Pr\left[Z_N\left(X\right) \le z\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx$$

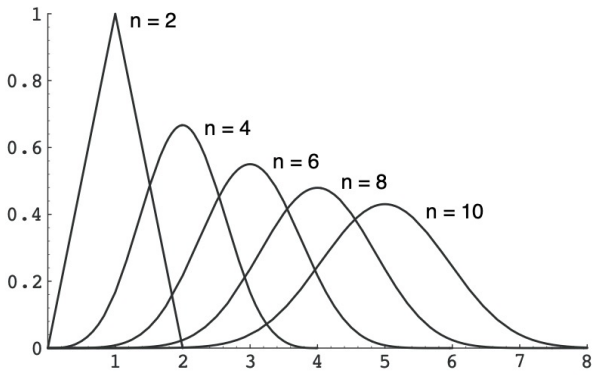
as N becomes large, the CDF of normalized and centered $Z_N\left(X\right)$ approaches Gaussian pdf

Sum of random variables: convergence to Gaussian

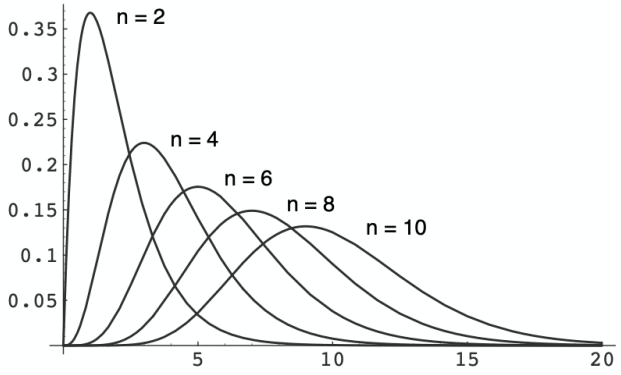
- Sum of independent random variables
  - Example: sum of two uniform independent random variables



Convolution of $N = n$ uniform pdfs



Convolution of $N = n$ exponential pdfs



- The pdf of sum converges to Gaussian pdf

Similarity between vectors

- We have seen the “similarity” between random variables expressed by covariance (via the second order moments) or independence (the level of pmf/pdf)

- Our next objectives:
  - To introduce a measure of similarity for deterministic vectors via a notion of metric and define:
    - Distance, inner product and cosine similarity
  - To extend it to random vectors:
    - Expected value of distance and inner product
    - Concentration properties (weak law of large numbers)
  - To extend the notion of similarity to distributions:
    - Divergence: KLD and f-div (Theme 2)
    - Mutual Information (Theme 2)

Vector norms

- Vector norms $\mathbf{x} = (x_1, x_2, \cdots, x_N)^T$

$\mathbf{x} = \mathbb{R}^N$

$\|\mathbf{x}\|_p$ — ℓ_p - norm or p- norm

$$\|\mathbf{x}\|_p = \left(|x_1|^p + |x_2|^p + \cdots |x_N|^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{N} |x_i|^p\right)^{\frac{1}{p}}$$

$\|\mathbf{x}\|_2$ — ℓ_2 - norm (Euclidian)

$$\|\mathbf{x}\|_2 = \left(|x_1|^2 + |x_2|^2 + \cdots |x_N|^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{N} |x_i|^2\right)^{\frac{1}{2}}$$

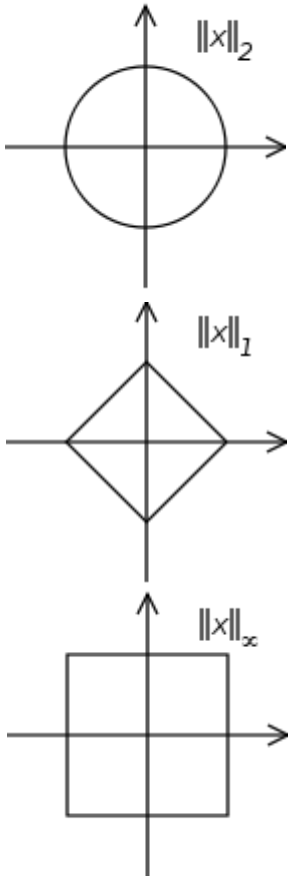
$\|\mathbf{x}\|_1$ — ℓ_1 - norm

$$\|\mathbf{x}\|_1 = \left(|x_1| + |x_2| + \cdots |x_N|\right) = \left(\sum_{i=1}^{N} |x_i|\right)$$

$\|\mathbf{x}\|_\infty$ — ℓ_∞ - norm or maximum norm

$$\|\mathbf{x}\|_\infty = \max\left\{|x_1|, |x_2|, \cdots, |x_N|\right\}$$

Unit-norm



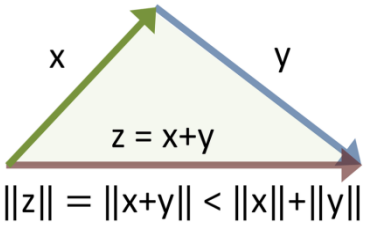
Vector norms

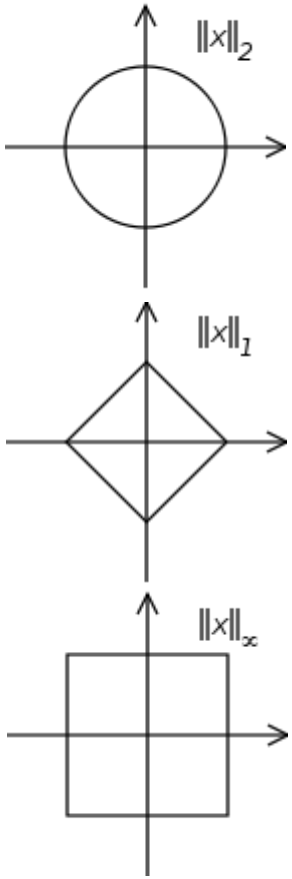
- Vector norms $\mathbf{x} = (x_1, x_2, \cdots, x_N)^T$

$\mathbf{x} = \mathbb{R}^N$

For all $p \geq 1$, the p-norms and maximum norm as defined above satisfy the properties of a "length function" (or norm), which are that:

- only the zero vector has zero length,
- the length of the vector is positive homogeneous with respect to multiplication by a scalar ($f(\alpha\mathbf{v}) = \alpha^k f(\mathbf{v})$)
- the length of the sum of two vectors is not larger than the sum of lengths of the vectors (triangle inequality).

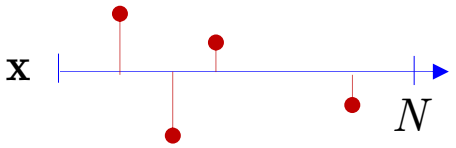




Vector ”norm” for sparse vectors

- The **number of nonzeros** is called the ℓ_0 –”norm” of $\mathbf{x}$

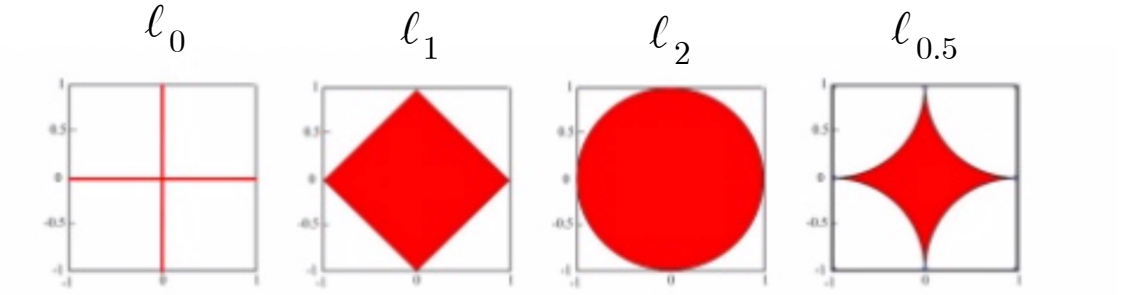
$$\left\|\mathbf{x}\right\|_0 \doteq \#\left\{i : x_i \neq 0\right\}$$



$$\left\|\mathbf{x}\right\|_0 = 4$$

$$\left\|\mathbf{x}\right\|_p = \left(\sum_{i=1}^{N} \left|x_i\right|^p\right)^{1/p}$$

$$\left\|\mathbf{x}\right\|_0 = \lim_{p \to 0} \left\|\mathbf{x}\right\|_p^p$$



Distances

Distances and similarity measures

A metric d is a function (mapping)

$$d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

that satisfies:

1. $d(x,y) \geq 0$ (*non-negative*)
2. $d(x,y) = 0$, iff $x = y$ (*identity of indiscernibles*)
3. $d(x,y) = d(y,x)$ (*symmetry*)
4. $d(x,z) \leq d(x,y) + d(y,z)$ (*triangle inequality*)

Distances

Distances and similarity measures

$\|\mathbf{x} - \mathbf{y}\|_2$ Euclidian distance between two vectors

$$d_2(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2 = \left(\sum_{i=1}^{N}(x_i - y_i)^2\right)^{\frac{1}{2}}$$

$\|\mathbf{x} - \mathbf{y}\|_1$ ℓ_1 - distance between two vectors

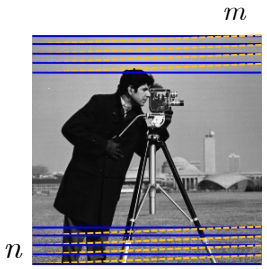
$$d_1(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_1 = \sum_{i=1}^{N}|x_i - y_i|$$

$\mathbf{x} \oplus \mathbf{y}$ Hamming distance between two vectors

$$d^H(\mathbf{x},\mathbf{y}) = \mathbf{x} \oplus \mathbf{y} = \sum_{i=1}^{N} 1\{x_i, y_i\}$$

with indicator function $1\{x_i, y_i\} = \begin{cases} 0, & \text{if } x_i = y_i \\ 1, & \text{if } x_i \neq y_i \end{cases}$

Distances



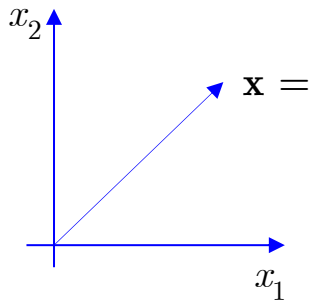
$$\begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} \quad N = n \times m$$

$\mathbf{x}_1$

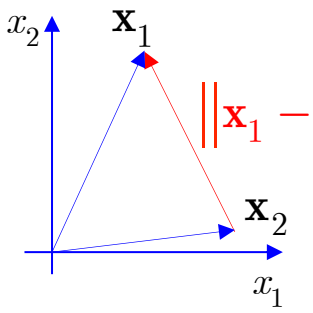
$\mathbf{x}_2$

$$d_2\left(\mathbf{x}_1, \mathbf{x}_2\right) = \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|_2 = \left(\sum_{i=1}^{N}\left(x_{1i} - x_{2i}\right)^2\right)^{\frac{1}{2}}$$

$n = 2$



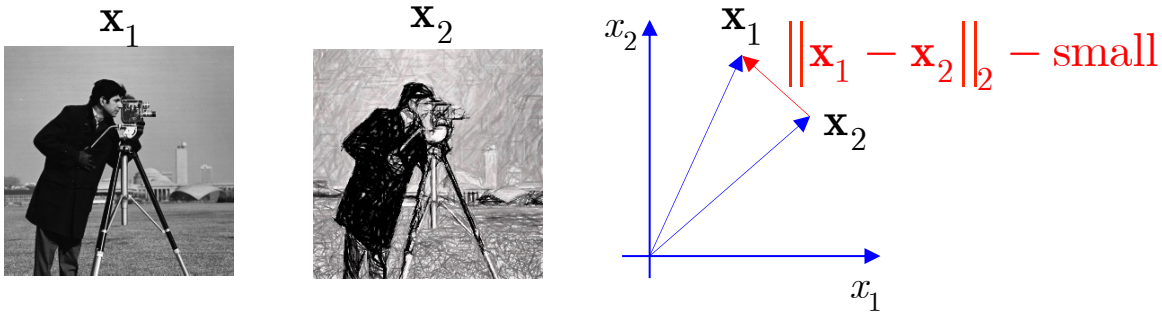
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$



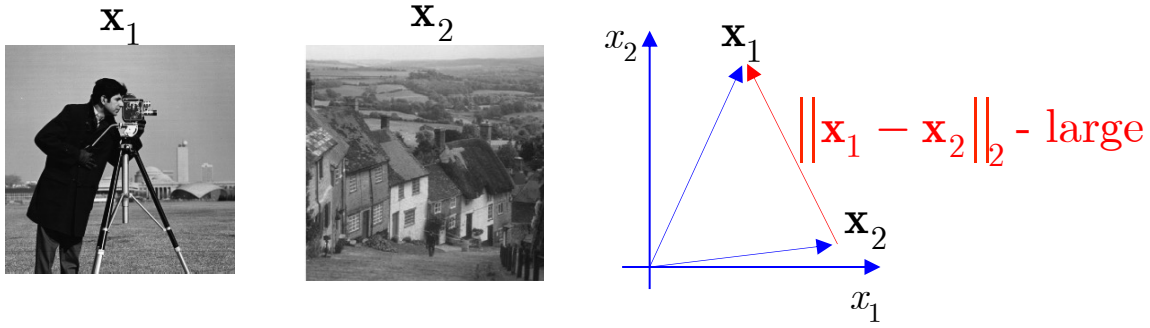
$\left\|\mathbf{x}_1 - \mathbf{x}_2\right\|_2$ - length

Distances

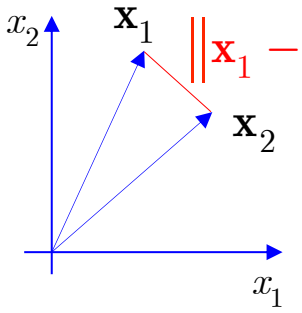
- Perceptually close images



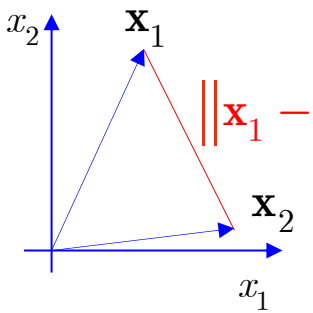
- Perceptually different images



Distances: link to inner product



$\|\mathbf{x}_1 - \mathbf{x}_2\|_2$ – small



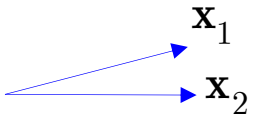
$\|\mathbf{x}_1 - \mathbf{x}_2\|_2$ - large

$$\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 = \|\mathbf{x}_1\|_2^2 - 2\underbrace{\mathbf{x}_1^T\mathbf{x}_2}_{\text{inner product}} + \|\mathbf{x}_2\|_2^2$$

Recall B: matrix operations

$$\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x} \in \mathbb{R} \triangleq \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} - & \mathbf{x} & - \end{bmatrix} \begin{bmatrix} | \\ \mathbf{y} \\ | \end{bmatrix} = \sum_{i=1}^{n} x_i y_i$$

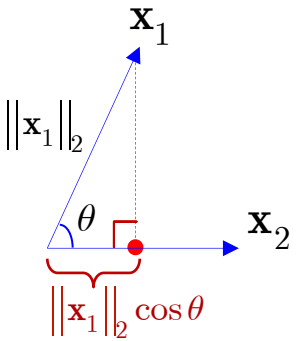
$\|\mathbf{x}_1 - \mathbf{x}_2\|_2$ – small $\Rightarrow$ $\mathbf{x}_1^T\mathbf{x}_2$ – large

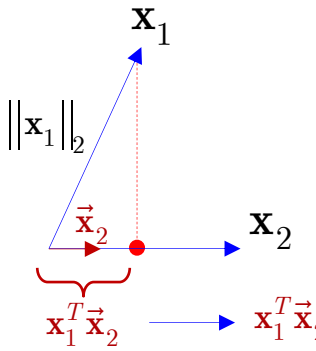


$\|\mathbf{x}_1 - \mathbf{x}_2\|_2$ – large $\Rightarrow$ $\mathbf{x}_1^T\mathbf{x}_2$ – small

Distances: link to cosine simiularity

- Cosine similarity





$$\vec{\mathbf{x}}_2 = \frac{\mathbf{x}_2}{\|\mathbf{x}_2\|_2} \text{ - unit vector/direction/}$$

$$\mathbf{x}_1^T\vec{\mathbf{x}}_2 = \mathbf{x}_1^T\left[\frac{\mathbf{x}_2}{\|\mathbf{x}_2\|_2}\right]$$

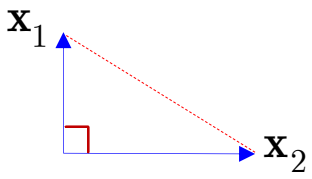
$$\|\mathbf{x}_1\|_2 \cos\theta = \mathbf{x}_1^T\vec{\mathbf{x}}_2 = \mathbf{x}_1^T\left[\frac{\mathbf{x}_2}{\|\mathbf{x}_2\|_2}\right]$$

$$\mathbf{x}_1^T\mathbf{x}_2 = \|\mathbf{x}_1\|_2\|\mathbf{x}_2\|_2 \cos\theta \qquad \Rightarrow \cos\theta = \frac{\mathbf{x}_1^T\mathbf{x}_2}{\|\mathbf{x}_1\|_2\|\mathbf{x}_2\|_2}$$

Distances: orthogonal and collinear vectors

$$\left\|\mathbf{x}_1-\mathbf{x}_2\right\|_2^2=\left\|\mathbf{x}_1\right\|_2^2+\left\|\mathbf{x}_2\right\|_2^2-2\mathbf{x}_1^T\mathbf{x}_2=\left\|\mathbf{x}_1\right\|_2^2+\left\|\mathbf{x}_2\right\|_2^2-2\cos\theta\left\|\mathbf{x}_1\right\|_2\left\|\mathbf{x}_2\right\|_2$$

- Orthogonal vectors

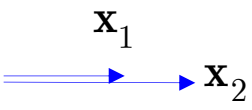


$$\mathbf{x}_1^T\mathbf{x}_2=0$$

$$\left\|\mathbf{x}_1\right\|_2\left\|\mathbf{x}_2\right\|_2\underbrace{\cos 90^0}_{=0}=0$$

$$\left\|\mathbf{x}_1-\mathbf{x}_2\right\|_2^2=\left\|\mathbf{x}_1\right\|_2^2+\left\|\mathbf{x}_2\right\|_2^2-2\underbrace{\mathbf{x}_1^T\mathbf{x}_2}_{=0}=\left\|\mathbf{x}_1\right\|_2^2+\left\|\mathbf{x}_2\right\|_2^2$$

- Collinear vectors



$$\mathbf{x}_1^T\mathbf{x}_2=\max$$

$$\left\|\mathbf{x}_1\right\|_2\left\|\mathbf{x}_2\right\|_2\underbrace{\cos 0^0}_{=1}=\left\|\mathbf{x}_1\right\|_2\left\|\mathbf{x}_2\right\|_2$$

$$\left\|\mathbf{x}_1-\mathbf{x}_2\right\|_2^2=\left\|\mathbf{x}_1\right\|_2^2+\left\|\mathbf{x}_2\right\|_2^2-2\left\|\mathbf{x}_1\right\|_2\left\|\mathbf{x}_2\right\|_2$$

$$\mathbf{x}_1=\mathbf{x}_2$$

$$\left\|\mathbf{x}_1\right\|_2=\left\|\mathbf{x}_2\right\|_2$$

$$\left\|\mathbf{x}_1-\mathbf{x}_2\right\|_2^2=0$$

Transformation of random variables

- General linear transform:
  - Random scalar variables
  - Random vectors
- Sum of random variables
  - Sum of independent random variables
  - Concentration inequalities
  - Weak Law of Large Numbers and CLT
  - Application to vector norms (see later after transformation of Gaussian random vectors)
- Transformation of Gaussian random vectors
  - Mean and covariance matrix after transformation
  - Pdf after transformation
  - Properties of transform
    - Decorrelation
    - Whitening

Second order models

- Plan
  - Introduce correlation and covariance matrices
  - Introduce uncorrelated and independent vectors
  - Introduce bivariate and multidimensional Gaussian pdfs
  - Investigate transformation of Gaussian random vectors:
    - pdf and moments

Second order stochastic models

- **Covariance and cross-covariance** Only for the second order statistics
  - Covariance matrix of vector $\mathbf{X} \in \mathbb{R}^N$ $\mathbf{K}_{\mathbf{xx}} \in \mathbb{R}^{N \times N}$

$$\mathbf{K}_{\mathbf{xx}} = \mathrm{Var}[\mathbf{X}] = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])^T\right] = \mathrm{E}\left[\mathbf{X}\mathbf{X}^T\right] - \mathrm{E}[\mathbf{X}]\mathrm{E}[\mathbf{X}]^T$$

  - Cross-covariance matrix between vectors $\mathbf{X} \in \mathbb{R}^N$ and $\mathbf{Y} \in \mathbb{R}^M$ $\mathbf{K}_{\mathbf{xy}} \in \mathbb{R}^{N \times M}$

$$\mathbf{K}_{\mathbf{xy}} = \mathrm{Cov}[\mathbf{X}, \mathbf{Y}] = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])^T\right] = \mathrm{E}\left[\mathbf{X}\mathbf{Y}^T\right] - \mathrm{E}[\mathbf{X}]\mathrm{E}[\mathbf{Y}]^T$$

  - Properties
    - The covariance matrix is a symmetric matrix $\mathbf{K}_{\mathbf{xx}}^T = \mathbf{K}_{\mathbf{xx}}$
    - The covariance matrix is a positive semidefinite matrix $\mathbf{a}^T \mathbf{K}_{\mathbf{xx}} \mathbf{a} \geq 0$
    - The transpose of cross-covariance matrix is: $\mathbf{K}_{\mathbf{xy}}^T = \mathbf{K}_{\mathbf{yx}}$

Second order stochastic models

- Example of covariance matrix

$$\mathbf{K}_{\mathbf{xx}} = \mathrm{Cov}[\mathbf{X},\mathbf{X}] = \mathrm{E}\left[(\mathbf{X}-\mathrm{E}[\mathbf{X}])(\mathbf{X}-\mathrm{E}[\mathbf{X}])^T\right] = \mathrm{E}\left[\mathbf{X}\mathbf{X}^T\right] - \mathrm{E}[\mathbf{X}]\mathrm{E}[\mathbf{X}]^T =$$

$$\begin{bmatrix} \mathrm{E}[(X_1-\mathrm{E}[X_1])(X_1-\mathrm{E}[X_1])] & \mathrm{E}[(X_1-\mathrm{E}[X_1])(X_2-\mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_1-\mathrm{E}[X_1])(X_N-\mathrm{E}[X_N])] \\ \mathrm{E}[(X_2-\mathrm{E}[X_2])(X_1-\mathrm{E}[X_1]) & \mathrm{E}[(X_2-\mathrm{E}[X_2])(X_2-\mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_2-\mathrm{E}[X_2])(X_N-\mathrm{E}[X_N])] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_N-\mathrm{E}[X_N])(X_1-\mathrm{E}[X_1])] & \mathrm{E}[(X_N-\mathrm{E}[X_N])(X_2-\mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_N-\mathrm{E}[X_N])(X_N-\mathrm{E}[X_N])] \end{bmatrix}$$

$$\mathbf{K}_{x_i x_j} = Cov\left[X_i, X_j\right] = \mathrm{E}\left[(X_i-\mathrm{E}[X_i])(X_j-\mathrm{E}[X_j])\right]$$

Second order stochastic models

- Example of cross-covariance matrix

$$\mathbf{K}_{\mathbf{xy}} = \mathrm{Cov}[\mathbf{X},\mathbf{Y}] = \mathrm{E}\left[(\mathbf{X}-\mathrm{E}[\mathbf{X}])(\mathbf{Y}-\mathrm{E}[\mathbf{Y}])^T\right] = \mathrm{E}\left[\mathbf{X}\mathbf{Y}^T\right] - \mathrm{E}[\mathbf{X}]\mathrm{E}[\mathbf{Y}]^T =$$

$$\begin{bmatrix} \mathrm{E}[(X_1-\mathrm{E}[X_1])(Y_1-\mathrm{E}[Y_1])] & \mathrm{E}[(X_1-\mathrm{E}[X_1])(Y_2-\mathrm{E}[Y_2])] & \cdots & \mathrm{E}[(X_1-\mathrm{E}[X_1])(Y_M-\mathrm{E}[Y_M])] \\ \mathrm{E}[(X_2-\mathrm{E}[X_2])(Y_1-\mathrm{E}[Y_1]) & \mathrm{E}[(X_2-\mathrm{E}[X_2])(Y_2-\mathrm{E}[Y_2])] & \cdots & \mathrm{E}[(X_2-\mathrm{E}[X_2])(Y_M-\mathrm{E}[Y_M])] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_N-\mathrm{E}[X_N])(Y_1-\mathrm{E}[Y_1])] & \mathrm{E}[(X_N-\mathrm{E}[X_N])(Y_2-\mathrm{E}[Y_2])] & \cdots & \mathrm{E}[(X_N-\mathrm{E}[X_N])(Y_M-\mathrm{E}[Y_M])] \end{bmatrix}$$

$$\mathbf{K}_{x_i y_j} = Cov[X_i, Y_j] = \mathrm{E}[(X_i-\mathrm{E}[X_i])(Y_j-\mathrm{E}[Y_j])]$$

Second order stochastic models

- **Correlation and cross-correlation** Only for the second order statistics
  - Correlation matrix of vector $\mathbf{X} \in \mathbb{R}^N$

$$\mathbf{R}_{\mathbf{xx}} = \mathrm{E}\left[\mathbf{X}\mathbf{X}^{\mathrm{T}}\right] \qquad \mathbf{R}_{\mathbf{xx}} \in \mathbb{R}^{N \times N}$$

  - Cross-correlation matrix between vectors $\mathbf{X} \in \mathbb{R}^N$ and $\mathbf{Y} \in \mathbb{R}^M$

$$\mathbf{R}_{\mathbf{xy}} = \mathrm{E}\left[\mathbf{X}\mathbf{Y}^{T}\right] \qquad \mathbf{R}_{\mathbf{xy}} \in \mathbb{R}^{N \times M}$$

  - Properties
    - The link between covariance and correlation

$$\mathbf{R}_{\mathbf{xx}} = \mathbf{K}_{\mathbf{xx}} + \mathrm{E}[\mathbf{X}]\mathrm{E}[\mathbf{X}]^{T}$$

    - The link between cross-covariance and cross-correlation

$$\mathbf{R}_{\mathbf{xy}} = \mathbf{K}_{\mathbf{xy}} + \mathrm{E}[\mathbf{X}]\mathrm{E}[\mathbf{Y}]^{T}$$

Summary: random vectors

- **Orthogonal** random vectors $\mathbf{X} \in \mathbb{R}^N$ and $\mathbf{Y} \in \mathbb{R}^M$

$$E\left[\mathbf{X}\mathbf{Y}^T\right] = \mathbf{0}_{N\times M}$$

- **Uncorrelated** random vectors $\mathbf{X} \in \mathbb{R}^N$ and $\mathbf{Y} \in \mathbb{R}^M$

$$E\left[\mathbf{X}\mathbf{Y}^T\right] = E\left[\mathbf{X}\right]E\left[\mathbf{Y}\right]^T = \underline{\mu}_X\underline{\mu}_Y^T$$

$$\Rightarrow Cov\left[\mathbf{X},\mathbf{Y}\right] = E\left[\mathbf{X}\mathbf{Y}^T\right] - \underline{\mu}_X\underline{\mu}_Y^T = E\left[\mathbf{X}\right]E\left[\mathbf{Y}\right]^T - \underline{\mu}_X\underline{\mu}_Y^T = \mathbf{0}$$

Only for the second order statistics

If $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated, then $\mathbf{X} - \underline{\mu}_X$ and $\mathbf{Y} - \underline{\mu}_Y$ are orthogonal.

- **Independent** random vectors $\mathbf{X} \in \mathbb{R}^N$ and $\mathbf{Y} \in \mathbb{R}^M$

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) \Rightarrow E\left[\mathbf{X}\mathbf{Y}^T\right] = E\left[\mathbf{X}\right]E\left[\mathbf{Y}^T\right] = \underline{\mu}_X\underline{\mu}_Y^T$$

$$E\left[\mathbf{X}\mathbf{Y}^T\right] = \int\int \mathbf{x}\mathbf{y}^T p(\mathbf{x},\mathbf{y})\,d\mathbf{x}\,d\mathbf{y} = \int \mathbf{x}p(\mathbf{x})\,d\mathbf{x}\int \mathbf{y}^T p(\mathbf{y})\,d\mathbf{y} = E\left[\mathbf{X}\right]E\left[\mathbf{Y}^T\right] = \underline{\mu}_X\underline{\mu}_Y^T$$

Second order models: Bivariate Gaussian

$$f(\mathbf{x}) = p(x_1, x_2) = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left(\frac{(x_1-\bar{x}_1)^2}{\sigma_{X_1}^2} + \frac{(x_2-\bar{x}_2)^2}{\sigma_{X_2}^2} - 2\frac{\rho(x_1-\bar{x}_1)(x_2-\bar{x}_2)}{\sigma_{X_1}\sigma_{X_2}}\right)\right]$$

$$\bar{\mathbf{x}} = \begin{bmatrix}\bar{x}_1 \\ \bar{x}_2\end{bmatrix}, \mathbf{K}_{\mathbf{xx}} = \begin{bmatrix}\sigma_{X_1}^2 & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma_{X_2}^2\end{bmatrix}, \quad \rho = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\mathrm{Var}(X_1)}}, \quad -1 \le \rho \le +1$$

$$\Rightarrow f(\mathbf{x}) = f(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 \left|\det \mathbf{K}_{\mathbf{xx}}\right|}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\bar{\mathbf{x}}\right)^T \mathbf{K}_{\mathbf{xx}}^{-1}\left(\mathbf{x}-\bar{\mathbf{x}}\right)\right]$$

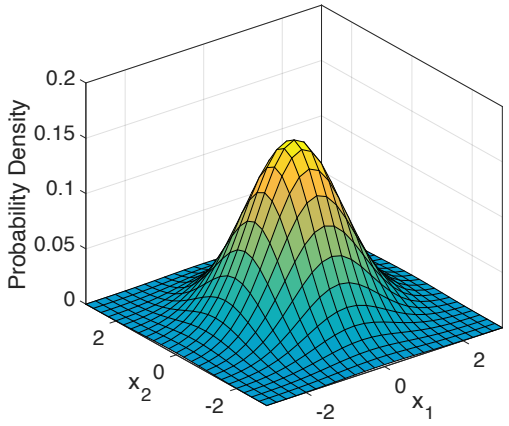
- Uncorrelated data $\rho = 0$: $\mathbf{K}_{\mathbf{xx}} = \begin{bmatrix}\sigma_{X_1}^2 & 0 \\ 0 & \sigma_{X_2}^2\end{bmatrix}$

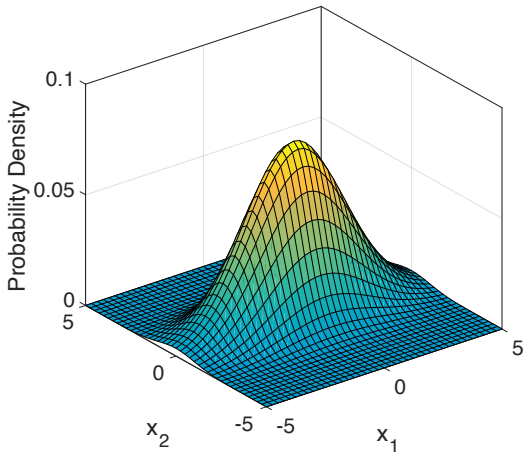
$$f(\mathbf{x}) = f(x_1, x_2) = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}} \exp\left[-\frac{1}{2}\left(\frac{(x_1-\bar{x}_1)^2}{\sigma_{X_1}^2} + \frac{(x_2-\bar{x}_2)^2}{\sigma_{X_2}^2}\right)\right]$$

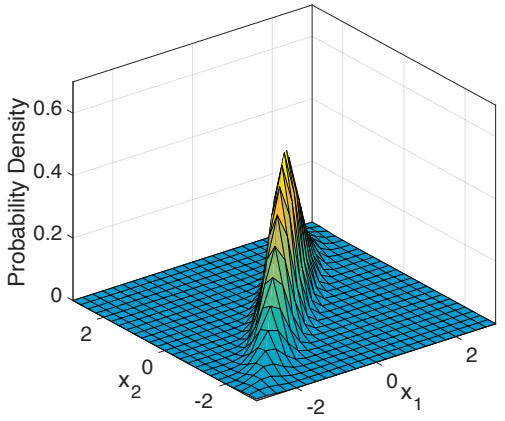
$$\Rightarrow f(\mathbf{x}) = f(x_1, x_2) = \underbrace{\frac{1}{\sqrt{2\pi\sigma_{X_1}^2}} \exp\left[-\frac{(x_1-\bar{x}_1)^2}{2\sigma_{X_1}^2}\right]}_{f(x_1)} \underbrace{\frac{1}{\sqrt{2\pi\sigma_{X_2}^2}} \exp\left[-\frac{(x_2-\bar{x}_2)^2}{2\sigma_{X_2}^2}\right]}_{f(x_2)}$$

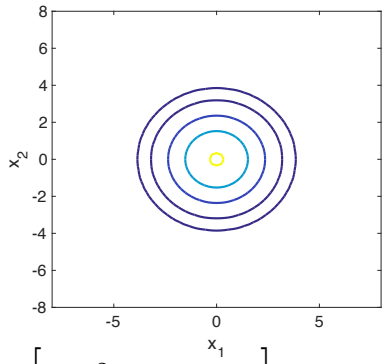
It implies independence ⟶

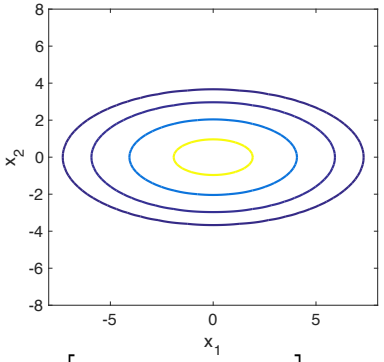
PCA on synthetic data: bivariate Gaussian

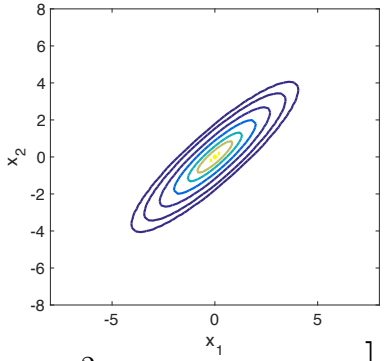












$$\begin{bmatrix} \sigma^2_{X_1} & 0 \\ 0 & \sigma^2_{X_2} \end{bmatrix} \sigma^2_{X_1} = \sigma^2_{X_2} = 1$$

$$\begin{bmatrix} \sigma^2_{X_1} & 0 \\ 0 & \sigma^2_{X_2} \end{bmatrix} \begin{matrix} \sigma^2_{X_1} = 4 \\ \sigma^2_{X_2} = 1 \end{matrix}$$

$$\begin{bmatrix} \sigma^2_{X_1} & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma^2_{X_2} \end{bmatrix} \begin{matrix} \sigma^2_{X_1} = 1 \\ \sigma^2_{X_2} = 1 \\ \rho = 0.9 \end{matrix}$$

Second order models: Multivariate Gaussian

$$f(\mathbf{x}) = f(x_1, x_2, \cdots, x_N) = \frac{1}{\sqrt{(2\pi)^N \left|\det \mathbf{K}_{\mathbf{xx}}\right|}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\bar{\mathbf{x}}\right)^T \mathbf{K}_{\mathbf{xx}}{}^{-1}\left(\mathbf{x}-\bar{\mathbf{x}}\right)\right]$$

$$\left(\mathbf{x}-\bar{\mathbf{x}}\right)^T \mathbf{K}_{\mathbf{xx}}^{-1}\left(\mathbf{x}-\bar{\mathbf{x}}\right) = \left\|\left(\mathbf{x}-\bar{\mathbf{x}}\right)\right\|^2_{\mathbf{K}_{\mathbf{xx}}^{-1}} = D_M \quad \text{- the Mahalanobis distance}$$

Notation: $\mathbf{X} \sim \mathcal{N}(\bar{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}})$

Remark: we talk about a random vector $\mathbf{X}$ generated from $\mathbf{X} \sim \mathcal{N}(\bar{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}})$

It should not be confused with the data matrix! It should be clear from the text.

For square matrix: $\mathbf{K}_{\mathbf{xx}} = \mathbf{U}\Sigma\mathbf{U}^{-1}$

- Eigenvectors $\mathbf{U} \in \underbrace{\mathbb{C}^{n\times n}}_{\text{generally complex}} = \begin{bmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \\ | & | & & | \end{bmatrix}$
- Eigenvalues $\sigma_1, \cdots, \sigma_n, \quad \underbrace{\sigma_i \in \mathbb{C}}_{\text{generally complex}} \quad \Sigma = diag\left(\sigma_1, \cdots, \sigma_n\right)$ or $\Sigma_{ii} = \sigma_i, \forall i$

Recall C: eigenvalue and SVD decompositions

Recall: for square and symmetric $\mathbf{K}_{\mathbf{xx}} = \mathbf{U}\Sigma\mathbf{U}^T$

Summary of notations

$\mathcal{N}(\mu, \sigma^2)$	Gaussian pdf with mean μ and variance σ^2
$X \sim \mathcal{N}(\bar{x}, \sigma_X^2)$	Scalar random variable with Gaussian pdf with mean $\bar{x}$ and variance σ_X^2
$\mathbf{X} \sim \mathcal{N}(\bar{\mathbf{x}}, \mathbf{C}_X)$	Vector random variable with Gaussian pdf with
$\mathbf{X} \sim \mathcal{N}(\bar{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}})$	mean vector $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{C}_X / \mathbf{K}_{\mathbf{xx}}$
$\mathbf{X} \sim \mathcal{N}(\bar{\mathbf{x}}, \sigma_X^2 \mathbf{I}_N)$	Vector RV with Gaussian pdf with mean vector $\bar{\mathbf{x}}$ and diagonal covariance matrix of size N with σ_X^2 on the main diagonal
$\mathcal{L}(\mu, \lambda)$	Laplacian pdf with mean μ and scale λ
$\mathcal{U}(a, b)$	Uniform pdf on interval [a,b]
$\mathcal{GGD}(\mu, \gamma, \lambda)$	Generalized Gaussian pdf with mean μ, shape parameter γ and scale λ
$\eta(\gamma, \lambda)$	Parameter of $\mathcal{GGD}$ pdf

Second order models: Multivariate Gaussian

- Generation of data with desired covariance matrix

$$\mathbf{X} \sim \mathcal{N}(\overline{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}})$$

$$\mathbf{K}_{\mathbf{xx}} = \mathbf{U}\Sigma\mathbf{U}^T = \mathbf{U}\Sigma^{1/2}\left(\mathbf{U}\Sigma^{1/2}\right)^T$$

$$\mathbf{X} \sim \mathcal{N}(\overline{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}}) \quad \Leftrightarrow \quad \mathbf{X} = \overline{\mathbf{x}} + \mathbf{U}\Sigma^{1/2}\mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$\Leftrightarrow \quad \mathbf{X} = \overline{\mathbf{x}} + \mathbf{U}\mathcal{N}(\mathbf{0}, \Sigma)$$

Linear transform of multivariate Gaussian vector

- **Linear transformation**:
  - Mean
  - Covariance matrix

  - Given an observation vector $\mathbf{x}$ generated from $\mathbf{X} \sim \mathcal{N}(\overline{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}})$
  - Apply a **generic linear transform** to this vector

$$\tilde{\mathbf{x}} = \mathbf{W}^T\mathbf{x} + \mathbf{b}$$

  - **Mean** of transformed vector

$$\tilde{\overline{\mathbf{x}}} = \mathbf{E}\left[\tilde{\mathbf{X}}\right] = \mathbf{E}\left[\mathbf{W}^T\mathbf{X} + \mathbf{b}\right] = \mathbf{W}^T\mathbf{E}\left[\mathbf{X}\right] + \mathbf{b} = \mathbf{W}^T\overline{\mathbf{x}} + \mathbf{b}$$

Remark: The Gaussian random vector under the linear (affine) transform will remain Gaussian but with modified mean and covariance.

Therefore, it suffices to find new mean and covariance matrix.

Linear transform of multivariate Gaussian vector

- **Linear transformation**:
  - Mean
  - Covariance matrix

    - **Covariance matrix** of transformed vector

$$\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \mathbf{E}\left[\left(\tilde{\mathbf{X}} - \mathbf{E}\left(\tilde{\mathbf{X}}\right)\right)\left(\tilde{\mathbf{X}} - \mathbf{E}\left(\tilde{\mathbf{X}}\right)\right)^T\right]$$

$$\tilde{\mathbf{X}} - \mathbf{E}\left(\tilde{\mathbf{X}}\right) = \mathbf{W}^T\mathbf{X} + \mathbf{b} - \mathbf{W}^T\overline{\mathbf{x}} - \mathbf{b} = \mathbf{W}^T\left(\mathbf{X} - \overline{\mathbf{x}}\right)$$

$$\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \mathbf{E}\left[\left(\mathbf{W}^T\left(\mathbf{X} - \overline{\mathbf{x}}\right)\right)\left(\mathbf{W}^T\left(\mathbf{X} - \overline{\mathbf{x}}\right)\right)^T\right] = \mathbf{E}\left[\mathbf{W}^T\left(\mathbf{X} - \overline{\mathbf{x}}\right)\left(\mathbf{X} - \overline{\mathbf{x}}\right)^T\mathbf{W}\right]$$

$$= \mathbf{W}^T\underbrace{\mathbf{E}\left[\left(\mathbf{X} - \overline{\mathbf{x}}\right)\left(\mathbf{X} - \overline{\mathbf{x}}\right)^T\right]}_{\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \mathbf{U}\Sigma\mathbf{U}^T}\mathbf{W} = \mathbf{W}^T\mathbf{C} = \mathbf{W}^T\mathbf{U}\Sigma\mathbf{U}^T\mathbf{W}$$

    - If $\mathbf{W}^T = \mathbf{U}^T$, i.e., the transform is chosen to be the decorrelation transform

$$\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \mathbf{W}^T\mathbf{U}\Sigma\mathbf{U}^T\mathbf{W} = \underbrace{\mathbf{U}^T\mathbf{U}}_{\mathbf{I}}\Sigma\underbrace{\mathbf{U}^T\mathbf{U}}_{\mathbf{I}} = \Sigma = diag\left\{\sigma_1, \cdots, \sigma_n\right\}$$

Linear transform of multivariate Gaussian vector

- **Decorrelation transform**:
  - Given an observation vector $\mathbf{x}$ generated from $\mathbf{X} \sim \mathcal{N}(\overline{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}})$
  - Apply the transform to the centered vector:

$$\tilde{\mathbf{x}} = \mathbf{U}^T\left(\mathbf{x} - \overline{\mathbf{x}}\right)$$

with $\mathbf{U}$ to be a eigenvector matrix of $\mathbf{K}_{\mathbf{xx}} = \mathbf{U}\Sigma\mathbf{U}^T$

  - The transformed data are also Gaussian $\tilde{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}})$ with the covariance matrix:

$$\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \mathbf{E}\left[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\right] = \mathbf{E}\left[\mathbf{U}^T\left(\mathbf{X} - \overline{\mathbf{x}}\right)\left(\mathbf{X} - \overline{\mathbf{x}}\right)^T\mathbf{U}\right] = \mathbf{U}^T\underbrace{\mathbf{E}\left[\left(\mathbf{X} - \overline{\mathbf{x}}\right)\left(\mathbf{X} - \overline{\mathbf{x}}\right)^T\right]}_{\mathbf{K}_{\mathbf{xx}}}\mathbf{U}$$

$$= \mathbf{U}^T\mathbf{K}_{\mathbf{xx}}\mathbf{U} = \underbrace{\mathbf{U}^T\mathbf{U}}_{\mathbf{I}}\Sigma\underbrace{\mathbf{U}^T\mathbf{U}}_{\mathbf{I}} = \Sigma$$

$$\mathbf{K}_{\mathbf{xx}} = \begin{bmatrix} \sigma^2_{X_1} & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma^2_{X_2} \end{bmatrix} \xrightarrow{\mathbf{U}^T} \mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$$

decorrelation

$$\rho \text{ any} \xrightarrow{\mathbf{U}^T} \tilde{\rho} = 0$$

Linear transform of multivariate Gaussian vector

- **Energy compaction (EC) property**:

$$\mathbf{K}_{\mathbf{xx}} = \begin{bmatrix} \sigma^2_{X_1} & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma^2_{X_2} \end{bmatrix} \xrightarrow{\mathbf{U}^T} \mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$$

All “energy” or information is only in the diagonal elements

$$\sigma^2_{X_1}, \sigma^2_{X_2} \text{ any} \xrightarrow{\mathbf{U}^T} \sigma_1 \geq \sigma_2$$

Decaying eigenvalues

Linear transform of multivariate Gaussian vector

- **Whitening property** :
  - Given an observation vector $\mathbf{x}$ generated from $\mathbf{X} \sim \mathcal{N}(\overline{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}})$
  - Apply whitening PCA to the centered vector:

$$\tilde{\mathbf{x}} = \Sigma^{-1/2}\mathbf{U}^T\left(\mathbf{x} - \overline{\mathbf{x}}\right)$$

with $\Sigma^{-1/2} = diag\left(1/\sqrt{\sigma_i}\right)$

  - The transformed data are also Gaussian with the covariance matrix: $\tilde{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$$\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \mathbf{E}\left[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\right] = \mathbf{E}\left[\Sigma^{-1/2}\mathbf{U}^T\left(\mathbf{X} - \overline{\mathbf{x}}\right)\left(\mathbf{X} - \overline{\mathbf{x}}\right)^T\mathbf{U}\Sigma^{-1/2}\right]$$

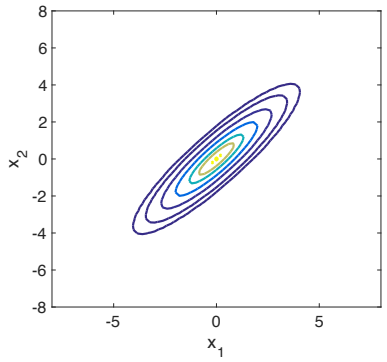
$$= \Sigma^{-1/2}\mathbf{U}^T\underbrace{\mathbf{E}\left[\left(\mathbf{X} - \overline{\mathbf{x}}\right)\left(\mathbf{X} - \overline{\mathbf{x}}\right)^T\right]}_{\mathbf{K}_{\mathbf{xx}}}\mathbf{U}\Sigma^{-1/2}$$

$$= \Sigma^{-1/2}\mathbf{U}^T\mathbf{K}_{\mathbf{xx}}\mathbf{U}\Sigma^{-1/2} = \underbrace{\Sigma^{-1/2}\mathbf{U}^T\mathbf{U}}_{\mathbf{I}}\Sigma\underbrace{\mathbf{U}^T\mathbf{U}\Sigma^{-1/2}}_{\mathbf{I}} = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = \mathbf{I}$$

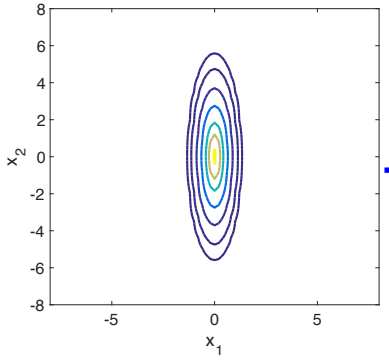
Demo

- **Matlab demo** $N = 2$

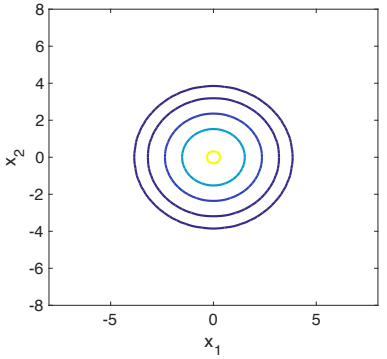
Original



Decorrelated



Whitened



$\mathbf{x}$

$\tilde{\mathbf{x}} = \mathbf{U}^T\left(\mathbf{x} - \overline{\mathbf{x}}\right)$

$\tilde{\mathbf{x}} = \Sigma^{-1/2}\mathbf{U}^T\left(\mathbf{x} - \overline{\mathbf{x}}\right)$

$\mathbf{K}_{\mathbf{xx}} = \mathbf{U}\Sigma\mathbf{U}^T$

$\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \Sigma$

$\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \mathbf{I}$

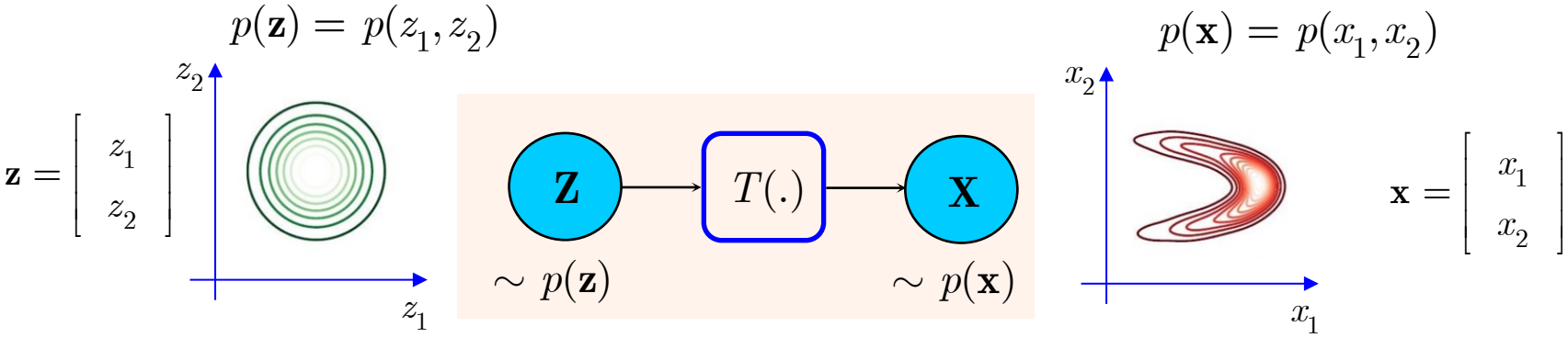
$\mathbf{K}_{\mathbf{xx}} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \rho = 0.9$

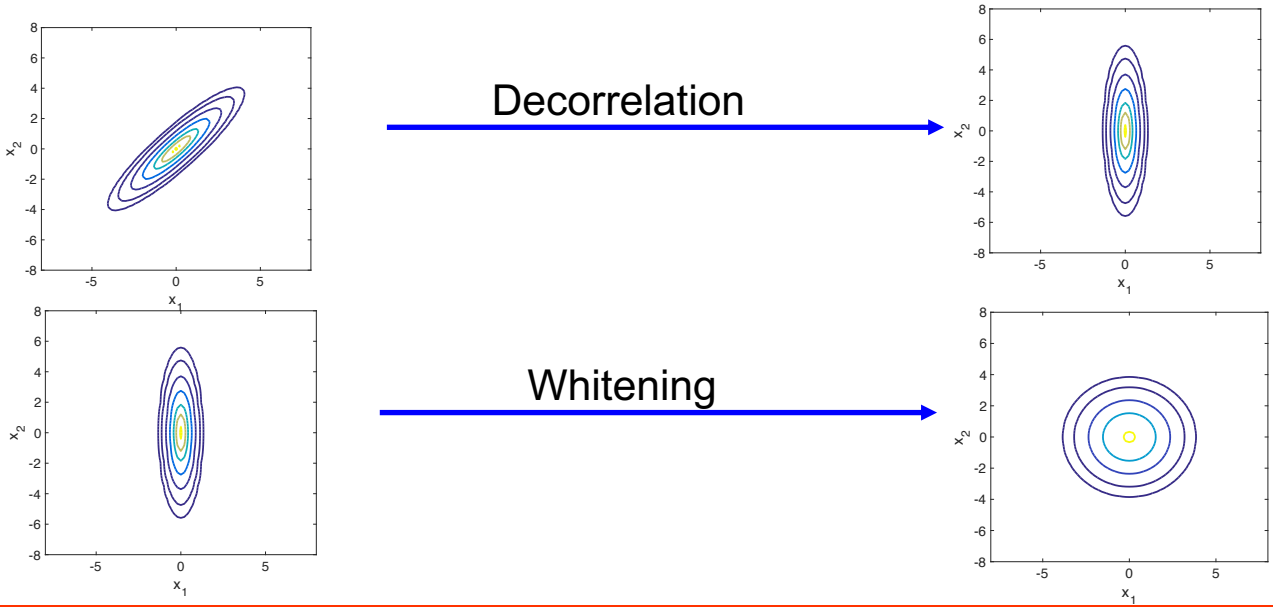
$\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \begin{pmatrix} 0.1 & 0 \\ 0 & 1.9 \end{pmatrix}$

$\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

Note: Matlab ordering

Recall: overall goal behind transformation





N-order Markov sequences

- **The N- order stationary Markov sequence:**

$$\mathbf{K}_{\mathbf{xx}} = \sigma_X^2 \begin{bmatrix} 1 & \rho & \rho^2 \cdots & \rho^{N-1} \\ \rho & 1 & \cdots & \rho^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \cdots & 1 \end{bmatrix} \quad \mathbf{X} \sim \mathcal{N}(\overline{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}}) \quad \Leftrightarrow \quad \mathbf{X} = \overline{\mathbf{x}} + \mathbf{U}\Sigma^{1/2}\mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$\Leftrightarrow \quad \mathbf{X} = \overline{\mathbf{x}} + \mathbf{U}\mathcal{N}(\mathbf{0}, \Sigma)$$

$$N = 4 \quad \rho = 0.95$$

$$\mathbf{K}_{\mathbf{xx}} = \sigma_X^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

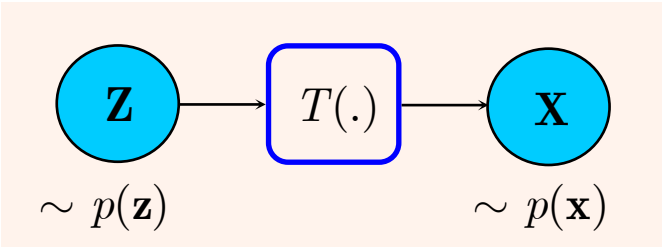
U = High frequency — Low frequency

0.2747	-0.5062	0.6516	0.4937
-0.6516	0.4937	0.2747	0.5062
0.6516	0.4937	-0.2747	0.5062
-0.2747	-0.5062	-0.6516	0.4937

D =

0.0300	0	0	0
0	0.0506	0	0
0	0	0.1627	0
0	0	0	3.7568

Linear transform: a link to affine transform



$$\mathbf{x} = \underbrace{\mathbf{W}^T}_{\mathbf{A}}\mathbf{z} = \mathbf{A}\mathbf{z}$$
$$\mathbf{z} = \mathbf{A}^{-1}\mathbf{x}$$

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{p_{\mathbf{z}}\left(T^{-1}(\mathbf{x})\right)}{\left|\det J_T(\mathbf{z})\right|} = \frac{p_{\mathbf{z}}\left(\mathbf{A}^{-1}\mathbf{x}\right)}{\left|\det J_T(\mathbf{z})\right|} = \frac{p_{\mathbf{z}}\left(\mathbf{A}^{-1}\mathbf{x}\right)}{\left|\det \mathbf{A}\right|}$$

- Consider $p(\mathbf{z}) = \mathcal{N}(\overline{\mathbf{z}}, \mathbf{K}_{\mathbf{zz}})$
- Find $p(\mathbf{x})$?

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{p_{\mathbf{z}}\left(\mathbf{A}^{-1}\mathbf{x}\right)}{\left|\det \mathbf{A}\right|} = \frac{1}{\left|\det \mathbf{A}\right|}\frac{1}{\sqrt{\left(2\pi\right)^N\left|\det \mathbf{K}_{\mathbf{zz}}\right|}}\exp\left[-\frac{1}{2}\left(\mathbf{A}^{-1}\mathbf{x}-\overline{\mathbf{z}}\right)^T\mathbf{K}_{\mathbf{zz}}{}^{-1}\left(\mathbf{A}^{-1}\mathbf{x}-\overline{\mathbf{z}}\right)\right]$$

Linear transform: a link to affine transform

- Simplification

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{p_{\mathbf{z}}\left(\mathbf{A}^{-1}\mathbf{x}\right)}{\left|\det \mathbf{A}\right|} = \frac{1}{\left|\det \mathbf{A}\right|} \frac{1}{\sqrt{\left(2\pi\right)^{N}\left|\det \mathbf{K}_{\mathbf{zz}}\right|}} \exp\left[-\frac{1}{2}\left(\mathbf{A}^{-1}\mathbf{x}-\overline{\mathbf{z}}\right)^{T} \mathbf{K}_{\mathbf{zz}}{}^{-1}\left(\mathbf{A}^{-1}\mathbf{x}-\overline{\mathbf{z}}\right)\right]$$

$$\left(\mathbf{A}^{-1}\mathbf{x}-\overline{\mathbf{z}}\right) = \mathbf{A}^{-1}\left(\mathbf{x}-\mathbf{A}\overline{\mathbf{z}}\right)$$

$$\left(\mathbf{A}^{-1}\mathbf{x}-\overline{\mathbf{z}}\right)^{T} = \left(\mathbf{A}^{-1}\left(\mathbf{x}-\mathbf{A}\overline{\mathbf{z}}\right)\right)^{T} = \left(\mathbf{x}-\mathbf{A}\overline{\mathbf{z}}\right)^{T}\mathbf{A}^{-1T}$$

$$\left(\mathbf{A}^{-1}\mathbf{x}-\overline{\mathbf{z}}\right)^{T} \mathbf{K}_{\mathbf{zz}}{}^{-1}\left(\mathbf{A}^{-1}\mathbf{x}-\overline{\mathbf{z}}\right) = \left(\mathbf{x}-\mathbf{A}\overline{\mathbf{z}}\right)^{T}\left(\mathbf{A}\mathbf{K}_{\mathbf{zz}}\mathbf{A}^{T}\right)^{-1}\left(\mathbf{x}-\mathbf{A}\overline{\mathbf{z}}\right)$$

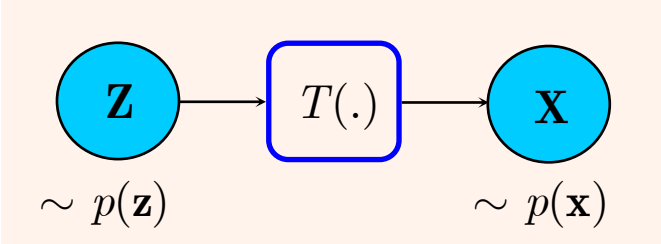
$$\mathbf{A}^{-1T}\mathbf{K}_{\mathbf{zz}}{}^{-1}\mathbf{A}^{-1} = \left(\mathbf{A}\mathbf{K}_{\mathbf{zz}}\mathbf{A}^{T}\right)^{-1} = \mathbf{K}_{\mathbf{xx}}^{-1}$$

$$\left|\det \mathbf{K}_{\mathbf{xx}}\right| = \left|\det \mathbf{A}\mathbf{K}_{\mathbf{zz}}\mathbf{A}^{T}\right| = \left|\det \mathbf{A}\right|\left|\det \mathbf{K}_{\mathbf{zz}}\right|\left|\det \mathbf{A}^{T}\right| = \left|\det \mathbf{A}\right|^{2}\left|\det \mathbf{K}_{\mathbf{zz}}\right|$$

$$\overline{\mathbf{x}} = \mathbf{A}\overline{\mathbf{z}}$$

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{\left(2\pi\right)^{N/2}\left|\det\left(\mathbf{K}_{\mathbf{xx}}\right)\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\overline{\mathbf{x}}\right)^{T} \mathbf{K}_{\mathbf{xx}}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}\right)\right]$$

Beyond second order models



- We considered generalized linear transform that targets the second order statistics

$$\mathbf{x} = \mathbf{W}^T\mathbf{z} + \mathbf{b}$$

- Possible extension is a non-linear transform

$$\mathbf{x} = \sigma\left(\mathbf{W}^T\mathbf{z} + \mathbf{b}\right)$$

  - Nested transforms = deep models

$$\mathbf{x} = \sigma_L\left(\mathbf{W}_L^T ... \sigma_1\left(\mathbf{W}_1^T\mathbf{z} + \mathbf{b}_1\right) + \mathbf{b}_L\right)$$

$$\Leftrightarrow \mathbf{x} = g_\theta\left(\mathbf{z}\right) \text{ with } \theta = \left(\mathbf{W}_1^T, ..., \mathbf{W}_L^T; \mathbf{b}_1, ..., \mathbf{b}_L\right)$$

Concentration properties and Gaussianity

- Concentration plays an important role for the analysis and design of systems

- Plan
  - Sum of jointly Gaussian random variables
    - As a special case of linear transform
  - Sum of squared random variables
    - The Euclidean norm of vector and its concentration
      - Mean
      - Variance
  - Hamming weight as a norm of binary sequences
      - Mean
      - Variance
  - Note on visualization

Linear transform: a link to affine transform

- A special case of linear transformation is

$$X = a_1 Z_1 + a_2 Z_2 + \cdots + a_N Z_N$$

where $Z_1, Z_2, \cdots, Z_N$ are jointly Gaussian

- X can be written

$$X = \begin{bmatrix} a_1 & \cdots & a_N \end{bmatrix} \begin{bmatrix} Z_1 \\ \vdots \\ Z_N \end{bmatrix} \triangleq \mathbf{AZ}$$

- X is a linear transformation of Gaussians that is also Gaussian

Linear transform: a link to affine transform

- Therefore, X is Gaussian with **mean**

$$E\left[X\right] = E\left[a_1 Z_1 + a_2 Z_2 + \cdots + a_N Z_N\right] = \sum_{i=1}^{N} a_i \underbrace{E\left[Z_i\right]}_{\mu_Z} = \mu_Z \sum_{i=1}^{N} a_i$$

where $Z_1, Z_2, \cdots, Z_N$ are jointly Gaussian

- **Variance**:

$$Var\left[X\right] = \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j Cov\left[Z_i, Z_j\right]$$

- If $Z_1, Z_2, \cdots, Z_N$ are independent Gaussian variables, i.e., $Cov\left[Z_i, Z_j\right] = 0$

- Then the variance of X reduces to

$$Var\left[X\right] = \sum_{i=1}^{N} a_i^2 Var\left[X_i\right]$$

Concentration properties and Gaussianity

- Concentration plays an important role for the analysis and design of systems

- Plan
  - Sum of jointly Gaussian random variables
    - As a special case of linear transform
  - Sum of squared random variables
    - The Euclidean norm of vector and its concentration
      - Mean
      - Variance
  - Hamming weight as a norm of binary sequences
      - Mean
      - Variance
  - Note on visualization

Concentration: ℓ_2 - norm

- **Concentration of norm around the mean**:
  - Given an i.i.d. unit variance random Gaussian vector $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - The Euclidean norm moments

$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ $\quad S_N(X) = \left\|\mathbf{X}\right\|_2^2 = X_1^2 + X_2^2 + \cdots X_N^2 \sim \chi^2$ with N degrees of freedom

$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I})$ $\quad S_N(X) = \left\|\mathbf{X}\right\|_2^2 = X_1^2 + X_2^2 + \cdots X_N^2 \sim Gamma$ with N degrees of freedom

  - **Mean**:

$$E\left[\left\|\mathbf{X}\right\|_2^2\right] = E\left[X_1^2 + \cdots X_N^2\right] = \sum_{i=1}^{N} \underbrace{E\left[X_i^2\right]}_{\sigma_X^2 = 1} = N$$

$N\sigma_X^2$

  - **Variance**:

$$Var\left[\left\|\mathbf{X}\right\|_2^2\right] = E\left[\left(\left\|\mathbf{X}\right\|_2^2\right)^2\right] - E\left[\left\|\mathbf{X}\right\|_2^2\right]^2$$

Concentration: ℓ_2 - norm

$$E\left[\left(\left\|\mathbf{X}\right\|_2^2\right)^2\right] = E\left[\left(\sum_{i=1}^{N} X_i^2\right)^2\right] = E\left[\sum_{i=1}^{N}\sum_{j=1}^{N} X_i^2 X_j^2\right] = \sum_{i=1}^{N}\sum_{j=1}^{N} E\left[X_i^2 X_j^2\right]$$

$$= \underbrace{\sum_{i=1}^{N} E\left[X_i^4\right]}_{3N\left(\sigma_X^2\right)^2} + 2\underbrace{\sum_{i=1}^{N-1}\sum_{j=i}^{N} E\left[X_i^2\right]E\left[X_j^2\right]}_{\text{sum of triangle matrix elements}} = \underbrace{(N-1)\frac{N}{2}}_{\sigma_X^2=1}\underbrace{(\sigma_{X_i}^2\sigma_{X_j}^2 + \sigma_{X_{ij}}^2)}_{=0}$$

$$3N + N(N-1) = N(N+2)$$

$$3N\left(\sigma_X^2\right)^2 + N(N-1)\left(\sigma_X^2\right)^2 = N(N+2)\left(\sigma_X^2\right)^2$$

$$Var\left[\left\|\mathbf{X}\right\|_2^2\right] = E\left[\left(\left\|\mathbf{X}\right\|_2^2\right)^2\right] - E\left[\left\|\mathbf{X}\right\|_2^2\right]^2 = N(N+2) - \left(N\right)^2 = 2N$$

$$2N\sigma_X^2$$

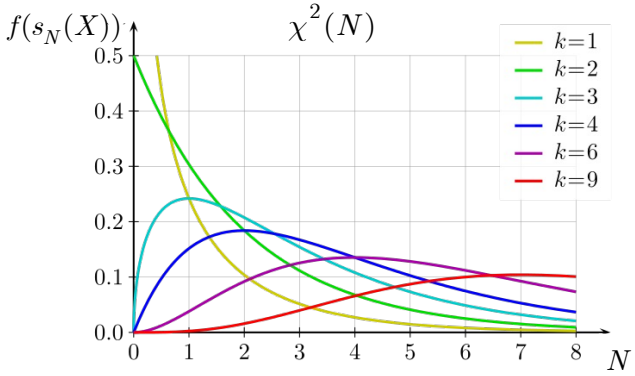
$$\frac{\text{Mean}}{\sqrt{\text{Variance}}} = \frac{E\left[\left\|\mathbf{X}\right\|_2^2\right]}{\sqrt{Var\left[\left\|\mathbf{X}\right\|_2^2\right]}} = \frac{N}{\sqrt{2N}} = \sqrt{\frac{N}{2}}$$

Mean growths faster than std

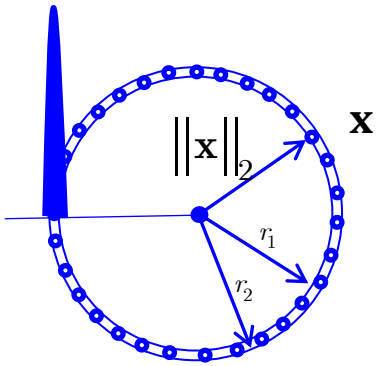
Summary: Euclidian norm concentration

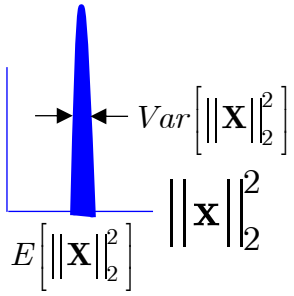
$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$\mathbf{x} = (x_1, \mathrm{L}\ , x_N) \qquad \|\mathbf{x}\|_2 = \sqrt{\underbrace{\sum_{i=1}^{N} x_i^2}_{x_1^2 + x_2^2 + \cdots + x_N^2}}$$

sum of N $\chi^2(N)$ unit variance R.V.s
$\Rightarrow$CLT$\Rightarrow$Gaussian



$$\left\{ \mathbf{x} : r_1 \le \|\mathbf{x}\|_2^2 \le r_2 \right\}$$





WLLN in action

$$\frac{\text{Mean}}{\sqrt{\text{Variance}}} = \frac{E\left[\|\mathbf{X}\|_2^2\right]}{\sqrt{Var\left[\|\mathbf{X}\|_2^2\right]}} = \frac{N}{\sqrt{2N}} = \sqrt{\frac{N}{2}}$$

Concentration: ℓ_2 - norm

- **Concentration of norm around the mean**:

- i.i.d. $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

- **Squared sum**:

$$S_N(X) = \left\|\mathbf{X}\right\|_2^2 \sim \chi^2(N)$$

- **Mean**:

$$Mean = E\left[\left\|\mathbf{X}\right\|_2^2\right] = N$$

- **Variance**:

$$Var\left\|\left\|\mathbf{X}\right\|_2^2\right\| = 2N$$

- i.i.d. $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2\mathbf{I})$

- **Squared sum**:

$$S_N(X) = \left\|\mathbf{X}\right\|_2^2 \sim Gamma(K, \theta)$$
$$K = N/2, \theta = 2\sigma_X^2$$

- **Mean**:

$$Mean = E\left[\left\|\mathbf{X}\right\|_2^2\right] = K\theta = \frac{N}{2}2\sigma_X^2 = N\sigma_X^2$$

- **Variance**:

$$Var\left\|\left\|\mathbf{X}\right\|_2^2\right\| = K\theta^2 = \frac{N}{2}\left(2\sigma_X^2\right)^2 = 2N\left(\sigma_X^2\right)^2$$

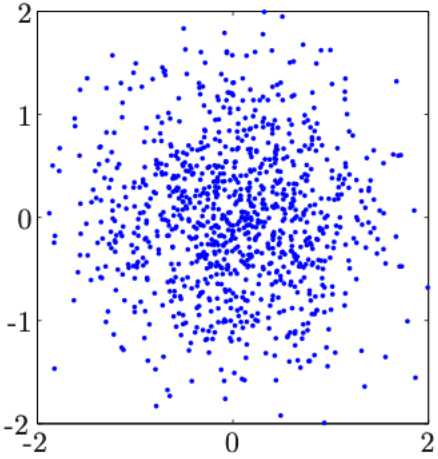
$$\frac{\text{Mean}}{\sqrt{\text{Variance}}} = \frac{E\left[\left\|\mathbf{X}\right\|_2^2\right]}{\sqrt{Var\left\|\left\|\mathbf{X}\right\|_2^2\right\|}} = \frac{N}{\sqrt{2N}} = \sqrt{\frac{N}{2}}$$

$$\frac{\text{Mean}}{\sqrt{\text{Variance}}} = \frac{E\left[\left\|\mathbf{X}\right\|_2^2\right]}{\sqrt{Var\left\|\left\|\mathbf{X}\right\|_2^2\right\|}} = \frac{N\sigma_X^2}{\sqrt{2N\left(\sigma_X^2\right)^2}} = \sqrt{\frac{N}{2}}$$

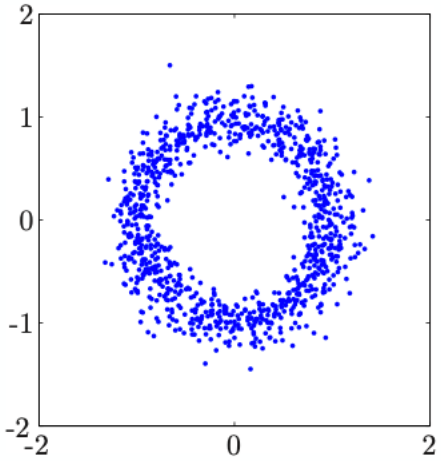
Concentration: ℓ_2- norm

- To visualize the effect of concentration in high dimensions assume $\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{N}\mathbf{I}\right)$

$\mathbb{R}^2$:



$\mathbb{R}^N$:



Here, only 2 first dim are visualized

- Vectors concentrate near the unit sphere $E\left[\left\|\mathbf{X}\right\|_2^2\right] = N\,Var\left[X_i\right] = N\frac{1}{N} = 1$
- Here, the variance decreases with the increase of dimension
  - Note: if the dimension increases, our model assumes the decrease of variance inversely proportionally to the dimension. Thus, with $N \to \infty$, we get random Gaussian vectors to be almost "deterministic" and near the unit sphere.

Concentration properties and Gaussianity

- Concentration plays an important role for the analysis and design of systems

- Plan
  - Sum of jointly Gaussian random variables
    - As a special case of linear transform
  - Sum of squared random variables
    - The Euclidean norm of vector and its concentration
      - Mean
      - Variance
  - Hamming weight as a norm of binary sequences
      - Mean
      - Variance
  - Note on visualization

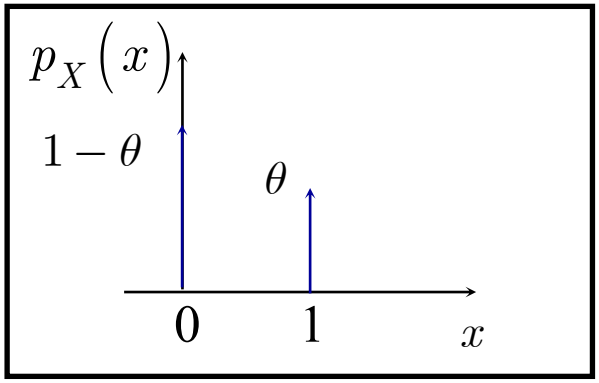
Recall: Binary or Bernouilli random variable

- Binary random variable takes value from the binary alphabet $X \in \mathcal{X} = \{0,1\}$

$$\theta \in [0,1] \quad p_X(x) = \begin{cases} 1-\theta, \, if \; x = 0, \\ \theta, \, if \; x = 1, \\ 0, \quad otherwise. \end{cases}$$

$$\Rightarrow \Pr[X_i = 1] = \theta, \Pr[X_i = 0] = 1-\theta$$

$$\Rightarrow \Pr[X = x] = \theta^x (1-\theta)^{1-x}$$



- **Moments**

$$E[X] = \theta \qquad Var[X] = \theta(1-\theta)$$

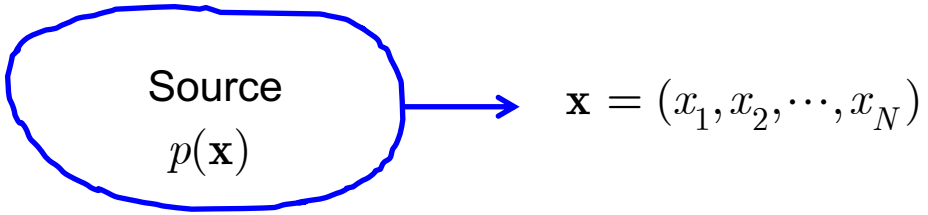
  - Proof

$$E[X] = 0 p_X(0) + 1 p_X(1) = 0(1-\theta) + 1\theta = \theta$$

$$Var[X] = E[X^2] - (E[X])^2 = (0^2(1-\theta) + 1^2\theta) - \theta^2 =$$
$$= \theta - \theta^2 = \theta(1-\theta)$$

Binomial pmf: as a sum of iid binary RVs

- Consider a source of iid binary random variables



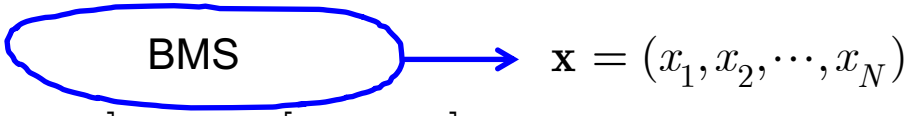
A binary memoryless source (BMS) produces a binary sequence $\mathbf{x} = (x_1, x_2, \cdots, x_N)$ with probability:

$$\Pr\left[\left(X_1, X_2, \cdots, X_N\right) = (x_1, x_2, \cdots, x_N)\right] = \prod_{i=1}^{N} \Pr\left[X_i = x_i\right]$$

with probabilities $\Pr\left[X_i = 1\right] = \theta, \Pr\left[X_i = 0\right] = 1 - \theta$ for $0 \le \theta < 1$ and $1 \le i \le N$.

Binomial pmf: as a sum of iid binary RVs

- Question: what is the probability to observe a sequence $\mathbf{x} = (x_1, x_2, \cdots, x_N)$?
- Given:



$$\Pr[X_i = 1] = \theta, \Pr[X_i = 0] = 1 - \theta$$

- Probability that the BMS generates a specific sequence $(x_1, x_2, \mathrm{L}\ , x_N)$ that contains ones (Hamming weight w) (and $(N - w)$ zeros) is:

$$\theta^w (1 - \theta)^{N-w}$$

- Probability that w occurs in a sequence $(x_1, x_2, \mathrm{L}\ , x_N)$ in any order is:

$$\Pr[W = w] = \binom{N}{w} \theta^w (1 - \theta)^{N-w}$$

in total there are $\binom{N}{w} = \frac{N!}{(N-w)!w!}$ sequences containing w ones.

Binomial pmf: as a sum of iid binary RVs

- Binomial pmf reflects the statistical behavior of sum of iid binary random variables
- It is also equivalent to the "norm" of binary vector = Hamming weight

$$\mathbf{X} = (X_1, X_2, \cdots, X_N) = (0, \underbrace{1}_{\text{with prob } \theta}, \cdots, \underbrace{1}_{\text{with prob } \theta})$$

Hamming weight $W = \sum_{i=1}^{N} X_i \qquad W \sim p_W(w) = \mathcal{B}(N, \theta)$

- **Moments**

$$E[W] = E\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} \underbrace{E[X_i]}_{\theta} = N\theta$$

$$Var[W] = \sum_{i=1}^{N} \underbrace{Var[X_i]}_{=\theta(1-\theta)} = N\theta(1-\theta)$$

Binomial pmf: as a sum of iid binary RVs

- Assume $\theta = 1/2$ and $\theta = 1/4$
  for $N = 32, 1024, 5024$

- Compute probability $\Pr\left[W = w\right]$ for $w = 0, 1, \cdots, N$ as binomial pmf

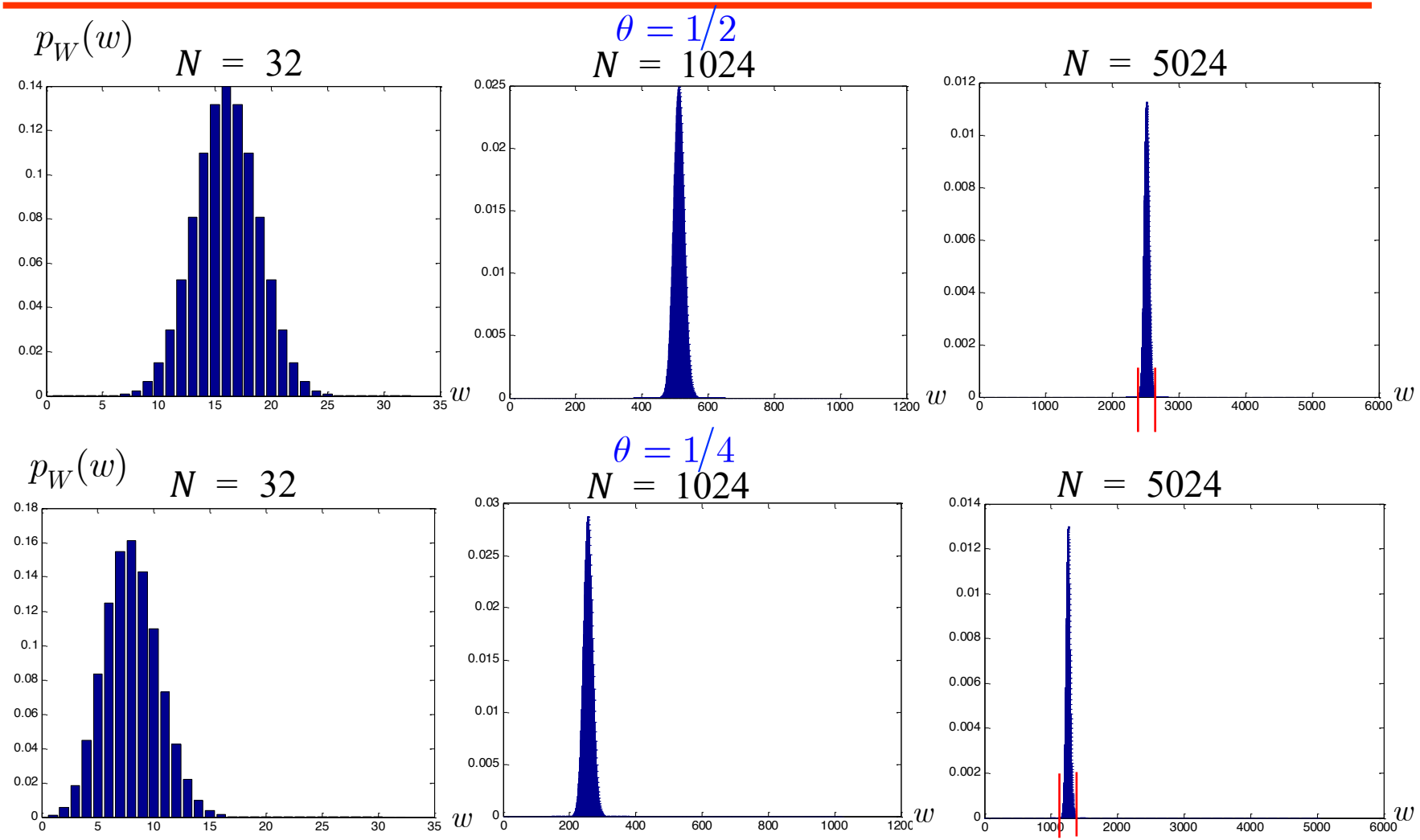
$$\Pr\left[W = w\right] = \underbrace{\binom{N}{w}}_{\#\text{ of N-length sequences with } w \text{ ones}} \times \underbrace{\theta^w \left(1-\theta\right)^{N-w}}_{\text{probability to observe } w \text{ ones in the sequence of length } N}$$

WLLN in action

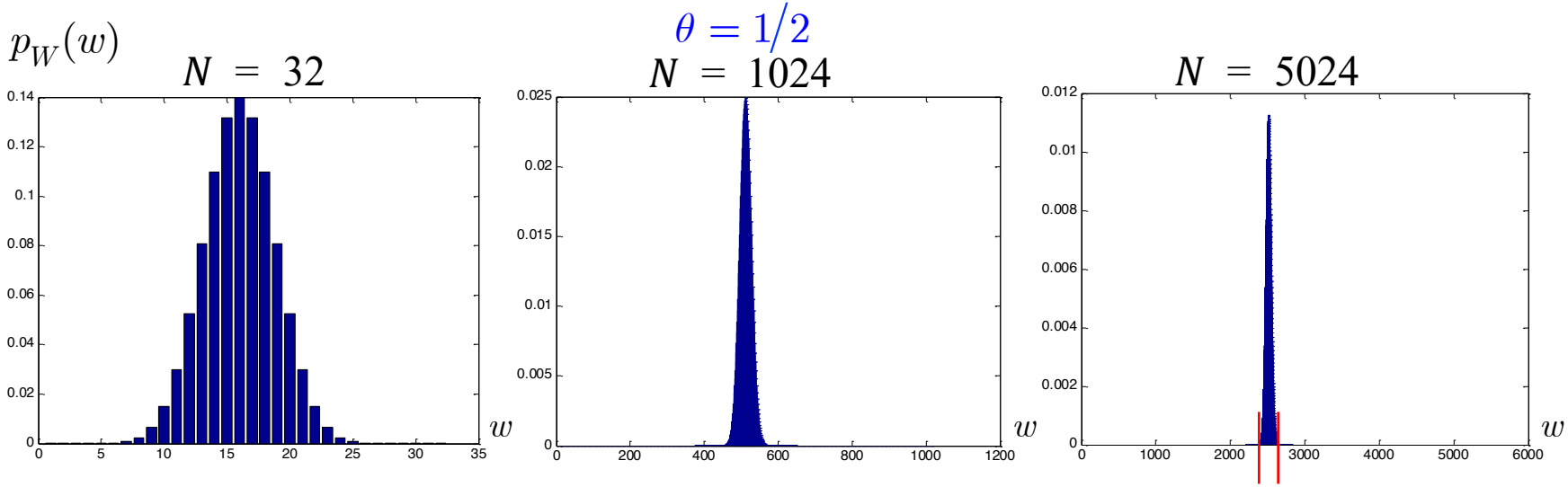
$$\frac{\text{Mean}}{\sqrt{\text{Variance}}} = \frac{E\left[W\right]}{\sqrt{Var\left[W\right]}} = \frac{N\theta}{\sqrt{N\theta(1-\theta)}} = \sqrt{\frac{N\theta}{(1-\theta)}}$$

Remarks: pay attention to mean, variance and probability concentration

Binomial pmf: as a sum of iid binary RVs



Binomial pmf: as a sum of iid binary RVs



WLLN in action

- **Remarks:** if the sequence length N increases:
  - the distribution of W peaks around $\theta N = 0.5N$ (mean $\Rightarrow$ WLLN)
  - the variance of distribution is $N\theta\left(1-\theta\right)$ $\Rightarrow$ pmf is peaky!

Concentration properties and Gaussianity

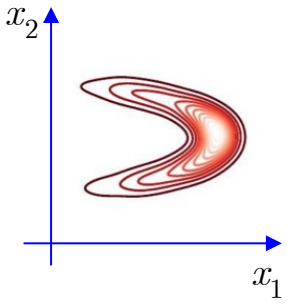
- Concentration plays an important role for the analysis and design of systems

- Plan
  - Sum of jointly Gaussian random variables
    - As a special case of linear transform
  - Sum of squared random variables
    - The Euclidean norm of vector and its concentration
      - Mean
      - Variance
  - Hamming weight as a norm of binary sequences
      - Mean
      - Variance
  - Note on visualization

Visualization of high dimensional data

- We can visualize data in 2D or 3D spaces

$$p(\mathbf{x}) = p(x_1, x_2)$$



$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Visualization of multidimensional data
  - Reduce dimensions to 2D by special transforms containing the most important data components

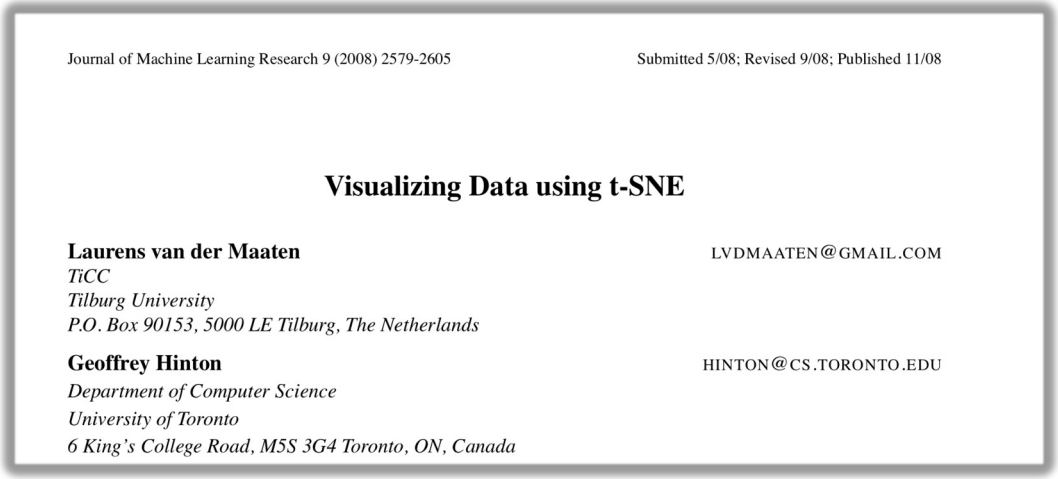
$$\tilde{\mathbf{x}} = \mathbf{W}^T \mathbf{x}$$

$$\tilde{\mathbf{x}} \in \mathbb{R}^2 \longleftarrow \mathbf{x} \in \mathbb{R}^N$$

  - Principle component analysis (PCA)
  - T-sne

Visualization of high dimensional data: T-sne

Stochastic Neighbor Embedding



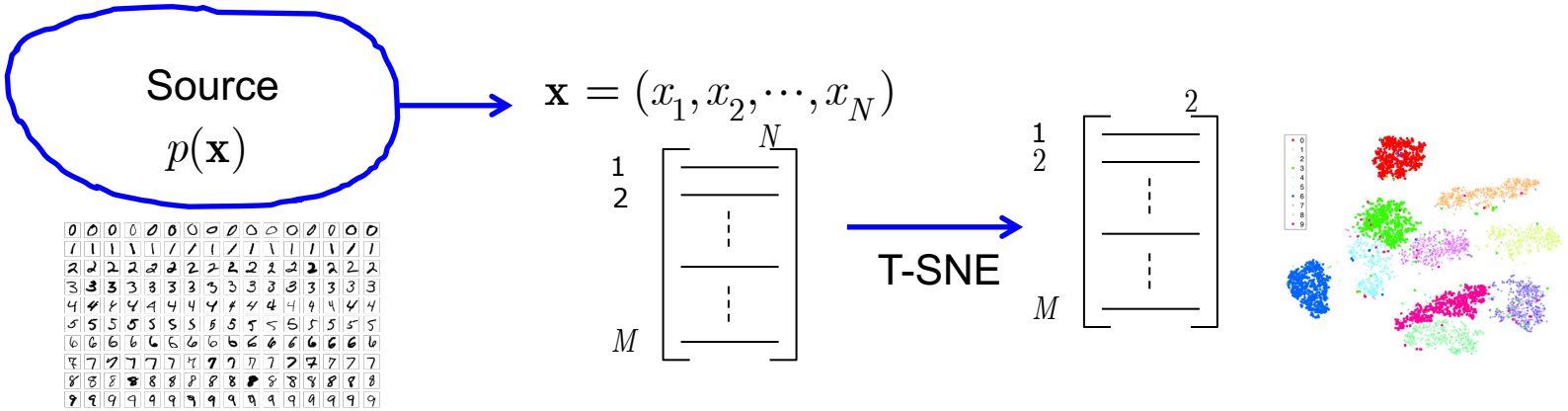
Journal of Machine Learning Research 9 (2008) 2579-2605 Submitted 5/08; Revised 9/08; Published 11/08

Visualizing Data using t-SNE

Laurens van der Maaten LVDMAATEN@GMAIL.COM
TiCC
Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Geoffrey Hinton HINTON@CS.TORONTO.EDU
Department of Computer Science
University of Toronto
6 King's College Road, M5S 3G4 Toronto, ON, Canada

http://jmlr.csail.mit.edu/papers/volume9/vandermaaten08a/vandermaaten08a.pdf



Appendix

- Recall A: Basic variance properties

Recall of variance and its properties

Definition (Variance of random variable):

$$Var\left[X\right] = \sigma_X^2 = E_{p(x)}\left[\left(X - \mu_X\right)^2\right]$$

where

$$\mu_X = E_{p(x)}\left[X\right] \qquad Var\left[X\right] = \sum_{x \in S_X} \left(x - \mu_X\right)^2 p_X\left(x\right)$$

- Standart deviation (Fr: écart-type) $\sigma_X = \sqrt{\sigma_X^2}$

Recall of variance and its properties

- Two forms of variance definition

$$Var[X] = E[X^2] - (E[X])^2 = E[X^2] - \mu_X{}^2$$ (Property 1)

  - Proof

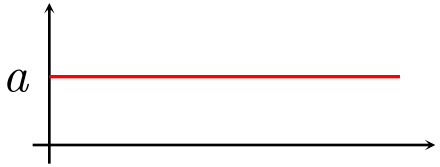
$$Var[X] = E\left[(X-\mu_X)^2\right] = \sum_{x\in S_X}(x-\mu_X)^2 p_X(x) =$$

$$= \sum_{x\in S_X} x^2 p_X(x) - 2\mu_X \underbrace{\sum_{x\in S_X} x p_X(x)}_{\mu_X} + \mu_X{}^2 \underbrace{\sum_{x\in S_X} p_X(x)}_{1} =$$

$$= E[X^2] - 2\mu_X{}^2 + \mu_X{}^2 = E[X^2] - \mu_X{}^2$$

- Variance of constant $Var[a] = 0$ (Property 2)
  - Proof

$$Var[a] = E\left[(a - E[a])^2\right] = E[0] = 0$$



Recall of variance and its properties

- Invariance under the mean shift. Assume b is a constant.

$$Var\left[X+b\right]=Var\left[X\right] \qquad \text{(Property 3)}$$

  - Proof

$$Var\left[X+b\right]=E\left[\left(\left(X+b\right)-E\left[X+b\right]\right)^2\right]=E\left[\left(X+b-E\left[X\right]-b\right)^2\right]$$
$$=E\left[\left(X-E\left[X\right]\right)^2\right]=Var\left[X\right]$$

- Scaling

$$Var\left[aX\right]=a^2Var\left[X\right] \qquad \text{(Property 4)}$$

  - Proof

$$Var\left[aX\right]=E\left[\left(aX-E\left[aX\right]\right)^2\right]=E\left[\left(aX-aE\left[X\right]\right)^2\right]$$
$$=E\left[a^2\left(X-E\left[X\right]\right)^2\right]=a^2Var\left[X\right]$$

Recall of variance and its properties

- Example of covaraince computation from joint distrubution

Example [edit]

Suppose that X and Y have the following joint probability mass function,[6] in which the six central cells give the discrete joint probabilities $f(x, y)$ of the six hypothetical realizations $(x, y) \in S = \{(1,1), (1,2), (1,3), (2,1), (2,2), (2,3)\}$:

			y		
	$f(x,y)$	1	2	3	$f_X(x)$
	1	1/4	1/4	0	1/2
x	2	0	1/4	1/4	1/2
	$f_Y(y)$	1/4	1/2	1/4	1

X can take on two values (1 and 2) while Y can take on three (1, 2, and 3). Their means are $\mu_X = 3/2$ and $\mu_Y = 2$. The population standard deviations of X and Y are $\sigma_X = 1/2$ and $\sigma_Y = \sqrt{2/3}$. Then:

$$\begin{aligned}
\operatorname{cov}(X,Y) = \sigma_{XY} &= \sum_{(x,y)\in S} f(x,y)(x-\mu_X)(y-\mu_Y) \\
&= \left(\frac{1}{4}\right)\left(1-\frac{3}{2}\right)(1-2) + \left(\frac{1}{4}\right)\left(1-\frac{3}{2}\right)(2-2) \\
&\quad + (0)\left(1-\frac{3}{2}\right)(3-2) + (0)\left(2-\frac{3}{2}\right)(1-2) \\
&\quad + \left(\frac{1}{4}\right)\left(2-\frac{3}{2}\right)(2-2) + \left(\frac{1}{4}\right)\left(2-\frac{3}{2}\right)(3-2) \\
&= \frac{1}{4}.
\end{aligned}$$

https://en.wikipedia.org/wiki/Covariance

Appendix

- Recall B: matrix operations

Recall of matrix operations: vector and matrix

- **Basic notations**
  - Vector

$$\mathbf{x} \in \mathbb{R}^n \triangleq \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{matrix} 1 \\ 2 \\ \\ \\ n \end{matrix} \overbrace{\bigcirc\!\!\bigcirc\!\!\bigcirc}^{\mathbf{x}}$$

  - Transposed vector

$$\mathbf{x}^T \in \mathbb{R}^{1\times n} \triangleq \begin{bmatrix} - & \mathbf{x} & - \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$$

  - Matrix

$$\mathbf{A} \in \mathbb{R}^{n\times m} = \begin{matrix} \\ \\ \\ n \end{matrix} \overbrace{\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}}^{m}$$

Recall of matrix operations: matrix forms

- **Representation of matrix**
  - Column form

$$\mathbf{A} = \begin{matrix} \\ n \end{matrix} \overset{\begin{matrix} | & | & & | \end{matrix}}{\begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_m \end{bmatrix}} \quad \overset{m}{}$$

  - Transposed column matrix

$$\mathbf{A}^T = \overset{n}{\begin{matrix} \\ m \end{matrix}} \begin{bmatrix} - & \mathbf{a}_1^T & - \\ - & \mathbf{a}_2^T & - \\ & \vdots & \\ - & \mathbf{a}_m^T & - \end{bmatrix}$$

  - Row form

$$\mathbf{A} = \begin{matrix} \\ n \end{matrix} \overset{m}{\begin{bmatrix} - & \vec{\mathbf{a}}_1^T & - \\ - & \vec{\mathbf{a}}_2^T & - \\ & \vdots & \\ - & \vec{\mathbf{a}}_n^T & - \end{bmatrix}}$$

  - Transposed row matrix

$$\mathbf{A}^T = \begin{matrix} \\ m \end{matrix} \overset{\begin{matrix} | & | & & | \end{matrix}}{\begin{bmatrix} \vec{\mathbf{a}}_1 & \vec{\mathbf{a}}_2 & \cdots & \vec{\mathbf{a}}_n \end{bmatrix}} \quad \overset{n}{}$$

Note: we always assume that the vector is represented as a column.

Recall of matrix operations: matrix multiplications

- **Vector-vector products**
  - Inner product

$$\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x} \in \mathbb{R} \triangleq \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} - & \mathbf{x} & - \end{bmatrix} \begin{bmatrix} | \\ \mathbf{y} \\ | \end{bmatrix} = \sum_{i=1}^{n} x_i y_i$$

  - Outer product

$$\mathbf{x}\mathbf{y}^T \in \mathbb{R}^{m\times n} \triangleq \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix}$$

$$\mathbf{x}\mathbf{y}^T = \begin{bmatrix} \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} y_1 & \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} y_2 & \cdots & \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} y_n \end{bmatrix}$$

Recall of matrix operations: matrix multiplications

- **Matrix-vector products**

$$\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m, \mathbf{x} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{m \times n} \qquad \mathbf{A} = \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} - & \vec{\mathbf{a}}_1^T & - \\ - & \vec{\mathbf{a}}_2^T & - \\ & \vdots & \\ - & \vec{\mathbf{a}}_m^T & - \end{bmatrix}$$

  - Inner form

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \begin{bmatrix} - & \vec{\mathbf{a}}_1^T & - \\ - & \vec{\mathbf{a}}_2^T & - \\ & \vdots & \\ - & \vec{\mathbf{a}}_m^T & - \end{bmatrix} \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} = \begin{bmatrix} \vec{\mathbf{a}}_1^T \mathbf{x} \\ \vec{\mathbf{a}}_2^T \mathbf{x} \\ \vdots \\ \vec{\mathbf{a}}_m^T \mathbf{x} \end{bmatrix}$$

  - Outer form

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \begin{bmatrix} | \\ \mathbf{a}_1 \\ | \end{bmatrix} x_1 + \cdots + \begin{bmatrix} | \\ \mathbf{a}_n \\ | \end{bmatrix} x_n$$

Recall of matrix operations: matrix multiplications

- **Matrix-matrix products**

$$\mathbf{AB},\ \mathbf{A} \in \mathbb{R}^{n\times m},\ \mathbf{B} \in \mathbb{R}^{m\times p}$$

$$\mathbf{A} = \begin{matrix} \\ n \end{matrix}\overset{m}{\begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_m \\ | & | & & | \end{bmatrix}} = \overset{m}{\begin{bmatrix} - & \vec{\mathbf{a}}_1^T & - \\ - & \vec{\mathbf{a}}_2^T & - \\ & \vdots & \\ - & \vec{\mathbf{a}}_n^T & - \end{bmatrix}}\ n$$

$$\mathbf{B} = \begin{matrix} \\ m \end{matrix}\overset{p}{\begin{bmatrix} | & | & & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_p \\ | & | & & | \end{bmatrix}} = \overset{p}{\begin{bmatrix} - & \vec{\mathbf{b}}_1^T & - \\ - & \vec{\mathbf{b}}_2^T & - \\ & \vdots & \\ - & \vec{\mathbf{b}}_m^T & - \end{bmatrix}}\ m$$

Recall of matrix operations: matrix multiplications

- **Matrix-matrix products**

$$\mathbf{AB},\ \mathbf{A} \in \mathbb{R}^{n \times m},\ \mathbf{B} \in \mathbb{R}^{m \times p}$$

  - Inner form

$$\mathbf{AB} = \begin{bmatrix} - & \vec{\mathbf{a}}_1^T & - \\ - & \vec{\mathbf{a}}_2^T & - \\ & \vdots & \\ - & \vec{\mathbf{a}}_n^T & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} \vec{\mathbf{a}}_1^T\mathbf{b}_1 & \vec{\mathbf{a}}_1^T\mathbf{b}_2 & \cdots & \vec{\mathbf{a}}_1^T\mathbf{b}_p \\ \vec{\mathbf{a}}_2^T\mathbf{b}_1 & \vec{\mathbf{a}}_2^T\mathbf{b}_2 & \cdots & \vec{\mathbf{a}}_2^T\mathbf{b}_p \\ \vdots & \vdots & \ddots & \vdots \\ \vec{\mathbf{a}}_n^T\mathbf{b}_1 & \vec{\mathbf{a}}_n^T\mathbf{b}_2 & \cdots & \vec{\mathbf{a}}_n^T\mathbf{b}_p \end{bmatrix}$$

$$= \begin{bmatrix} | & | & & | \\ \mathbf{Ab}_1 & \mathbf{Ab}_2 & \cdots & \mathbf{Ab}_p \\ | & | & & | \end{bmatrix}$$

Recall of matrix operations: matrix multiplications

- **Matrix-matrix products**

$$\mathbf{AB},\ \mathbf{A} \in \mathbb{R}^{n\times m},\ \mathbf{B} \in \mathbb{R}^{m\times p}$$

  - Outer form

$$\mathbf{AB} = \underset{n}{\begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_m \\ | & | & & | \end{bmatrix}} \overset{p}{\begin{bmatrix} - & \vec{\mathbf{b}}_1^T & - \\ - & \vec{\mathbf{b}}_2^T & - \\ & \vdots & \\ - & \vec{\mathbf{b}}_m^T & - \end{bmatrix}}$$

$$= \underbrace{\underset{n}{\begin{bmatrix} | \\ \mathbf{a}_1 \\ | \end{bmatrix}} \overset{p}{\begin{bmatrix} - & \vec{\mathbf{b}}_1^T & - \end{bmatrix}}}_{n\times p} + \underbrace{\underset{n}{\begin{bmatrix} | \\ \mathbf{a}_2 \\ | \end{bmatrix}} \overset{p}{\begin{bmatrix} - & \vec{\mathbf{b}}_2^T & - \end{bmatrix}}}_{n\times p} + \cdots + \underbrace{\underset{n}{\begin{bmatrix} | \\ \mathbf{a}_m \\ | \end{bmatrix}} \overset{p}{\begin{bmatrix} - & \vec{\mathbf{b}}_m^T & - \end{bmatrix}}}_{n\times p}$$

Applications: sample covariance matrix

- **Sample covariance matrix (assuming zero-mean)**

$$\mathbf{C} = \frac{1}{N}\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{n\times n}, \mathbf{X} \in \mathbb{R}^{n\times N}$$

$$\mathbf{X} = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \\ | & | & & | \end{bmatrix} \quad n$$

Assume zero mean vectors

$$\mathbf{C} = \frac{1}{N}\begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \\ | & | & & | \end{bmatrix}\begin{bmatrix} - & \mathbf{x}_1^T & - \\ - & \mathbf{x}_2^T & - \\ & \vdots & \\ - & \mathbf{x}_N^T & - \end{bmatrix}$$

- Outer form

$$\mathbf{C} = \frac{1}{N}\begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \\ | & | & & | \end{bmatrix}\begin{bmatrix} - & \mathbf{x}_1^T & - \\ - & \mathbf{x}_2^T & - \\ & \vdots & \\ - & \mathbf{x}_N^T & - \end{bmatrix} = \frac{1}{N}\left[\mathbf{x}_1\mathbf{x}_1^T + \mathbf{x}_2\mathbf{x}_2^T + \cdots + \mathbf{x}_N\mathbf{x}_N^T\right]$$

$$= \underbrace{\begin{bmatrix} | \\ \mathbf{x}_1 \\ | \end{bmatrix}\begin{bmatrix} - & \mathbf{x}_1^T & - \end{bmatrix}}_{n\times n} + \underbrace{\begin{bmatrix} | \\ \mathbf{x}_2 \\ | \end{bmatrix}\begin{bmatrix} - & \mathbf{x}_2^T & - \end{bmatrix}}_{n\times n} + \cdots + \underbrace{\begin{bmatrix} | \\ \mathbf{x}_N \\ | \end{bmatrix}\begin{bmatrix} - & \mathbf{x}_N^T & - \end{bmatrix}}_{n\times n}$$

Applications: Gram matrix

- **Gram matrix**

$$\mathbf{G} = \mathbf{X}^T\mathbf{X} \in \mathbb{R}^{N\times N}, \mathbf{X} \in \mathbb{R}^{n\times N}$$

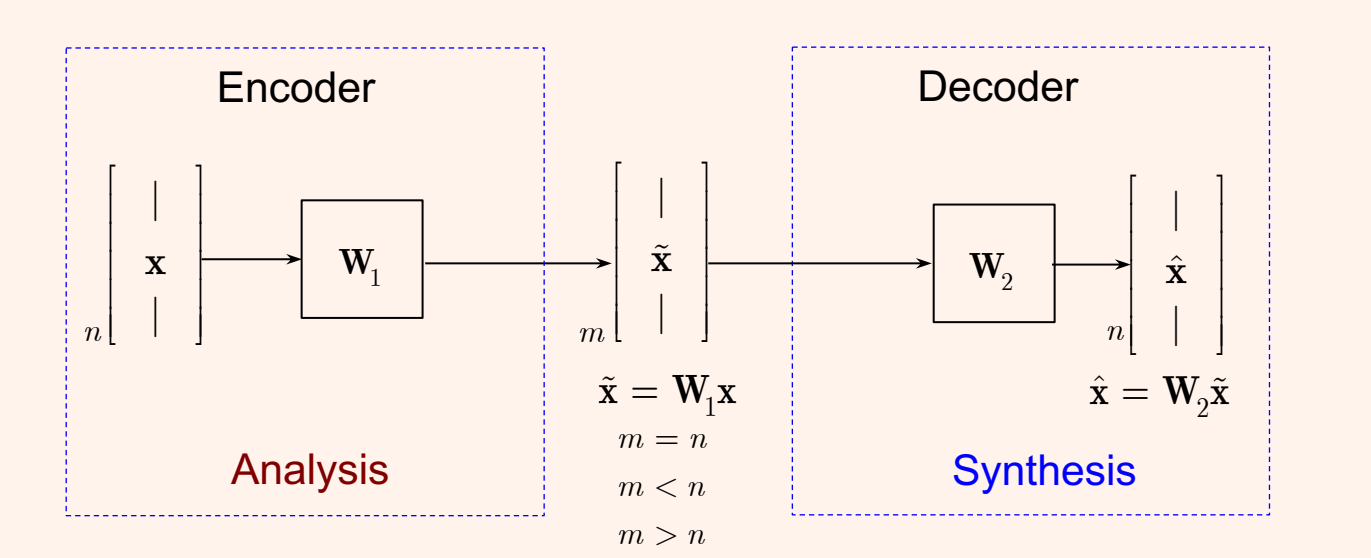
$$\mathbf{X} = \begin{matrix} n \\ \end{matrix}\begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \\ | & | & & | \end{bmatrix}$$

  - Inner form

$$\mathbf{C} = \begin{matrix} N \\ \end{matrix}\begin{bmatrix} - & \mathbf{x}_1^T & - \\ - & \mathbf{x}_2^T & - \\ & \vdots & \\ - & \mathbf{x}_N^T & - \end{bmatrix}\begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \\ | & | & & | \end{bmatrix}^N = \begin{bmatrix} \mathbf{x}_1^T\mathbf{x}_1 & \mathbf{x}_1^T\mathbf{x}_2 & \cdots & \mathbf{x}_1^T\mathbf{x}_N \\ \mathbf{x}_2^T\mathbf{x}_1 & \mathbf{x}_2^T\mathbf{x}_2 & \cdots & \mathbf{x}_2^T\mathbf{x}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_N^T\mathbf{x}_1 & \mathbf{x}_N^T\mathbf{x}_2 & \cdots & \mathbf{x}_N^T\mathbf{x}_N \end{bmatrix}$$

Applications: encoder-decoder notation

- **Encoder-decoder notations**



$$\hat{\mathbf{x}} = \mathbf{W}_2\tilde{\mathbf{x}} = \mathbf{W}_2\mathbf{W}_1\mathbf{x}$$

- Reconstruction (if no other constraints on properties of $\tilde{\mathbf{x}}$) for given $\mathbf{W}_1$ and $\mathbf{W}_2$

$$\left\|\mathbf{x} - \hat{\mathbf{x}}\right\|_2^2 = \left\|\mathbf{x} - \mathbf{W}_2\mathbf{W}_1\mathbf{x}\right\|_2^2 = \left\|\left(\mathbf{I} - \mathbf{W}_2\mathbf{W}_1\right)\mathbf{x}\right\|_2^2 \quad \Rightarrow \quad \text{For perfect reconstruction}$$

$$\left(\mathbf{I} - \mathbf{W}_2\mathbf{W}_1\right) = \mathbf{0} \rightarrow \quad \mathbf{I} = \mathbf{W}_2\mathbf{W}_1$$

Applications: encoder-decoder notation

- **Encoder-decoder notations**
  - Reconstruction

$$\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_2^2=\left\|\mathbf{x}-\mathbf{W}_2\mathbf{W}_1\mathbf{x}\right\|_2^2=\left\|\left(\mathbf{I}-\mathbf{W}_2\mathbf{W}_1\right)\mathbf{x}\right\|_2^2 \quad \Rightarrow \quad \mathbf{W}_2\mathbf{W}_1=\mathbf{I}$$

  - Special case $m=n,\ \mathbf{W}_2=\mathbf{A}\in\mathbb{R}^{n\times n}=\begin{bmatrix}| & | & & |\\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n\\ | & | & & |\end{bmatrix}$ - orthonormal
    Orthonormal matrices

$$\mathbf{A}^T\mathbf{A}=\mathbf{A}\mathbf{A}^T=\mathbf{I}$$
$$\mathbf{a}_i^T\mathbf{a}_i=1,\quad \mathbf{a}_i^T\mathbf{a}_j=0$$

$$\mathbf{W}_1=\mathbf{A}^T\in\mathbb{R}^{n\times n}=\begin{bmatrix}- & \mathbf{a}_1^T & -\\ - & \mathbf{a}_2^T & -\\ & \vdots & \\ - & \mathbf{a}_n^T & -\end{bmatrix}$$

$$\hat{\mathbf{x}}=\mathbf{W}_2\mathbf{W}_1\mathbf{x}=\underbrace{\mathbf{A}\mathbf{A}^T}_{\mathbf{I}}\mathbf{x}=\mathbf{x}$$

  - Otherwise, when $m\neq n$, a special care should be taken about the selection of encoder-decoder pairs. We will see more details in pseudo-inverse part.

Appendix

- Recall C: eigen value decomposition and SVD

Eigen (value) decomposition

- **Summary (if you are experienced)**

  - $\mathbf{X}$ is a square real matrix

    $$\mathbf{X} \in \mathbb{R}^{n \times n} = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ | & | & & | \end{bmatrix}$$

    - Eigenvectors

      $$\mathbf{U} \in \underbrace{\mathbb{C}^{n \times n}}_{\text{generally complex}} = \begin{bmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \\ | & | & & | \end{bmatrix} \quad \mathbf{U}\mathbf{U}^{-1} = \mathbf{I}$$

    - Eigenvalues $\sigma_1, \cdots, \sigma_n$, $\underbrace{\sigma_i \in \mathbb{C}}_{\text{generally complex}}$

      $$\Sigma = diag\left(\sigma_1, \cdots, \sigma_n\right) \text{ or } \Sigma_{ii} = \sigma_i, \forall i$$

  - Eigen decomposition

    $$\mathbf{X} = \mathbf{U}\Sigma\mathbf{U}^{-1}$$

Eigen (value) decomposition

- **Summary (if you are experienced)**
  - $\mathbf{X}$ is a (square) symmetric real matrix
    - Eigenvectors are orthonormal

$$\mathbf{U}^{-1} = \mathbf{U}^T$$

    - Eigenvalues $\sigma_1, \cdots, \sigma_n$ are real $\sigma_i \in \mathbb{R}$

$$\mathbf{X} \in \mathbb{R}^{n \times n} = \begin{bmatrix} a & \times & & \cdots & \circ \\ \times & b & & & \\ & & c & & \\ \vdots & & & \ddots & \\ \circ & & & & z \end{bmatrix}$$

  - Eigen decomposition of symmetric real matrix

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{U}^T$$

Matlab

```
[U,D] = eig(A)
```

Eigen (value) decomposition (you can skip it)

- **Friendly introduction**
  - $\mathbf{X}$ is any square real matrix
    - **Definition**: the eigenvector $\mathbf{u}$ is a vector that satisfies:

$$\mathbf{X}\mathbf{u} = \sigma\mathbf{u}, \mathbf{u} \neq \mathbf{0} \Rightarrow \left(\mathbf{X} - \sigma\mathbf{I}\right)\mathbf{u} = \mathbf{0}$$

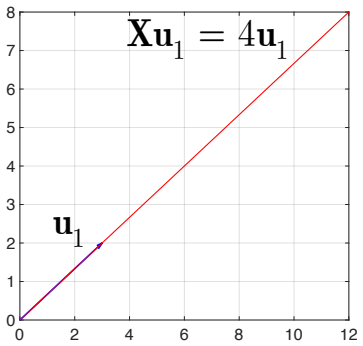
Example

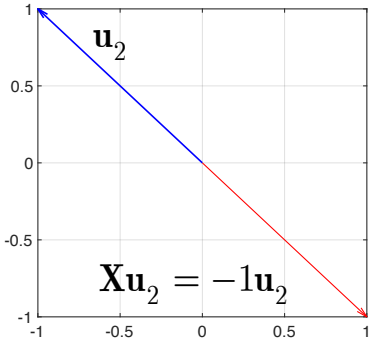
$$\mathbf{X} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$$ has the eigenvectors $$\mathbf{u}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \text{ with } \sigma_1{=}4$$

$$\mathbf{u}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \text{ with } \sigma_2{=}\text{-}1$$

$$\mathbf{X}\mathbf{u}_1 = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \underbrace{4}_{\sigma_1}\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix}$$

$$\mathbf{X}\mathbf{u}_2 = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \underbrace{-1}_{\sigma_2}\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$





Eigen (value) decomposition (you can skip it)

- **Friendly introduction**
  - In most applications, the eigenvectors are normalized

$$\mathbf{u}_i^T \mathbf{u}_i = 1$$

Example

$$\mathbf{u}_1 = \begin{bmatrix} 0.83 \\ 0.55 \end{bmatrix}, \text{with } \sigma_1 = 4$$

$$\mathbf{u}_2 = \begin{bmatrix} -0.71 \\ 0.71 \end{bmatrix}, \text{with } \sigma_2 = -1$$

Matlab

```
>> X=[2 3; 2 1]
X=
     2     3
     2     1

>> [U,D] = eig(X)
U =
    0.8321   -0.7071
    0.5547    0.7071
D =
     4     0
     0    -1
```

Eigen (value) decomposition (you can skip it)

- **Friendly introduction**
  - **Definition**: matrix form

$$\mathbf{X}\mathbf{u}_i = \sigma_i \mathbf{u}_i, \mathbf{u}_i \neq \mathbf{0}, \forall i \Rightarrow \mathbf{X}\mathbf{U} = \mathbf{U}\Sigma$$

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{U}^{-1}$$

Example: to check it out that indeed $\mathbf{X} = \mathbf{U}\Sigma\mathbf{U}^{-1}$

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{U}^{-1} = \begin{bmatrix} 3 & -1 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 4 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} 0.2 & 0.2 \\ -0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$$

Matlab

```
>> res=U*D*inv(U)
res =
    2.0000    3.0000
    2.0000    1.0000
```

Eigen (value) decomposition (you can skip it)

- **Friendly introduction**
  - **Definition 1**: positive (semi-)definite matrices
    - A matrix is said to be positive semi-definite when it can be obtained as the product of a matrix by its transpose

$$\mathbf{C} = \mathbf{X}\mathbf{X}^T$$

    - This implies that the positive semi-definite matrix is always symmetric.

  - **Definition 2 (alternative)**: positive (semi-)definite matrices
    - We can investigate the sign of eigenvalues of symmetric $\mathbf{X}$
    - Consider a decomposition $\mathbf{a}^T\mathbf{X}\mathbf{a}$ for any non-zero vector $\mathbf{a}$

$$\mathbf{a}^T\mathbf{X}\mathbf{a} = \underbrace{\mathbf{a}^T\mathbf{U}}_{\mathbf{y}^T=\mathbf{a}^T\mathbf{U}} \Sigma \underbrace{\mathbf{U}^T\mathbf{a}}_{\mathbf{y}=\mathbf{U}^T\mathbf{a}} = \sum_{i=1}^{n} \sigma_i \underbrace{y_i^2}_{\geq 0}$$

Since y_i^2 is always positive, the sign of the above quadratic from is fully determined by the sign of eigenvalues.

Eigen (value) decomposition (you can skip it)

- **Friendly introduction**
  - **Definition 2**: positive (semi-)definite matrices
    - If all $\sigma_i \geq 0$, then the matrix $\mathbf{X}$ is said to be **positive semi-definite**
    - If all $\sigma_i > 0$, then the matrix $\mathbf{X}$ is said to be **positive definite**
    - If all $\sigma_i \leq 0$, then the matrix $\mathbf{X}$ is said to be **negative semi-definite**
    - If all $\sigma_i < 0$, then the matrix $\mathbf{X}$ is said to be **negative definite**

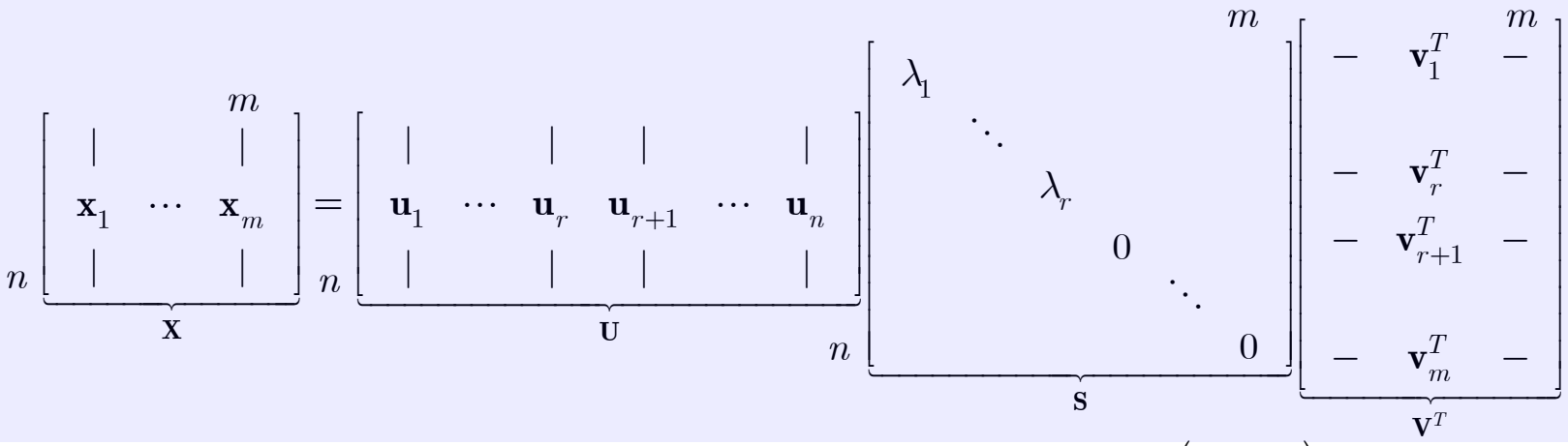
Eigen (value) decomposition (you can skip it)

- **Friendly introduction**
  - Properties of positive (semi-)definite matrices
    - Its eigenvalues are always positive or null, i.e., $\sigma_i \geq 0$
    - The eigenvectors are pairwise orthonormal, if their eigenvalues are different
    - The eigenvectors are real
    - Since all eigenvectors are orthonormal, i.e., $\mathbf{U}^{-1} = \mathbf{U}^T$
  - Rank of matrix:
    - the number of non-zero eigenvalues of the matrix and $r \leq \min\{n, m\}$
    - If rank is equal to the matrix dimension, it is called a full rank matrix
    - Otherwise, it is called rank-deficient, singular or multicolinear matrix
    - To have a full rank, all vectors should be linearly independent, i.e., none on them can be created as a linear combination of remaining ones

Singular value decomposition (SVD)

- **Fundamentals of SVD (can be skipped, if familiar)**

- For any $\mathbf{X} \in \mathbb{R}^{n \times m}$, there exists a factorization $\underbrace{\mathbf{X}}_{n \times m} = \underbrace{\mathbf{U}}_{n \times n} \underbrace{\mathbf{S}}_{n \times m} \underbrace{\mathbf{V}}_{m \times m}^{T}$



- $\mathbf{U}$ is an $n \times n$ orthonormal matrix consisting of $\mathbf{U} = \mathbf{evec}\left(\mathbf{X}\mathbf{X}^T\right)$
- $\mathbf{V}$ is an $m \times m$ orthonormal matrix consisting of $\mathbf{V} = \mathbf{evec}\left(\mathbf{X}^T\mathbf{X}\right)$
- $\mathbf{S}$ is an $n \times m$ diagonal matrix containing $r \leq \min\{n, m\}$ singular values $\lambda_i \geq 0$ on the main diagonal and $\mathbf{S}^2 = \mathbf{eval}\left(\mathbf{X}^T\mathbf{X}\right) = \mathbf{eval}\left(\mathbf{X}\mathbf{X}^T\right)$

Singular value decomposition (SVD)

- Properties of SVD:
  - $\underbrace{\mathbf{X}\mathbf{X}^T}_{n\times n}$ and $\underbrace{\mathbf{X}^T\mathbf{X}}_{m\times m}$ are square symmetric matrices that can be decomped as:

EVD: $\mathbf{X}\mathbf{X}^T = \mathbf{U}\Sigma\mathbf{U}^T$ where $\Sigma = \mathbf{eval}\left(\mathbf{X}\mathbf{X}^T\right)$

$\mathbf{X}^T\mathbf{X} = \mathbf{V}\Sigma'\mathbf{V}^T$ $\Sigma' = \mathbf{eval}\left(\mathbf{X}^T\mathbf{X}\right)$

Σ and Σ' have the same non-zero diagonal entries but in the different order

Validate: $\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T\left(\mathbf{U}\mathbf{S}\mathbf{V}^T\right)^T = \mathbf{U}\mathbf{S}\underbrace{\mathbf{V}^T\mathbf{V}}_{\mathbf{I}}\mathbf{S}^T\mathbf{U}^T = \mathbf{U}\underbrace{\mathbf{S}\mathbf{S}^T}_{\mathbf{S}^2=\Sigma,\ \lambda_i^2=\sigma_i}\mathbf{U}^T = \mathbf{U}\Sigma\mathbf{U}^T$

Singular value decomposition (SVD)

- **Demo example**

```
X =
     3     1     1
    -1     3     1


>> [U,S,V] = svd(A)
U =
   -0.7071   -0.7071
   -0.7071    0.7071
S =
    3.4641         0         0
         0    3.1623         0
V =
   -0.4082   -0.8944   -0.1826
   -0.8165    0.4472   -0.3651
   -0.4082    0.0000    0.9129
```

λ_i

Singular value decomposition (SVD)

- **Fundamentals of SVD**
  - **Demo example**

```
X =
     3     1     1
    -1     3     1
```

```
>> [U,D] = eig(X*X')
U =
   -0.7071    0.7071
    0.7071    0.7071
D =
   10.0000         0
         0   12.0000
```

1. Order
2. Squared

$\mathbf{S}^2=\Sigma, \lambda_i^2=\sigma_i$

```
>> [V,D] = eig(X'*X)
V =
    0.1826    0.8944    0.4082
    0.3651   -0.4472    0.8165
   -0.9129         0    0.4082
D =
         0         0         0
         0   10.0000         0
         0         0   12.0000
```

Symmetric

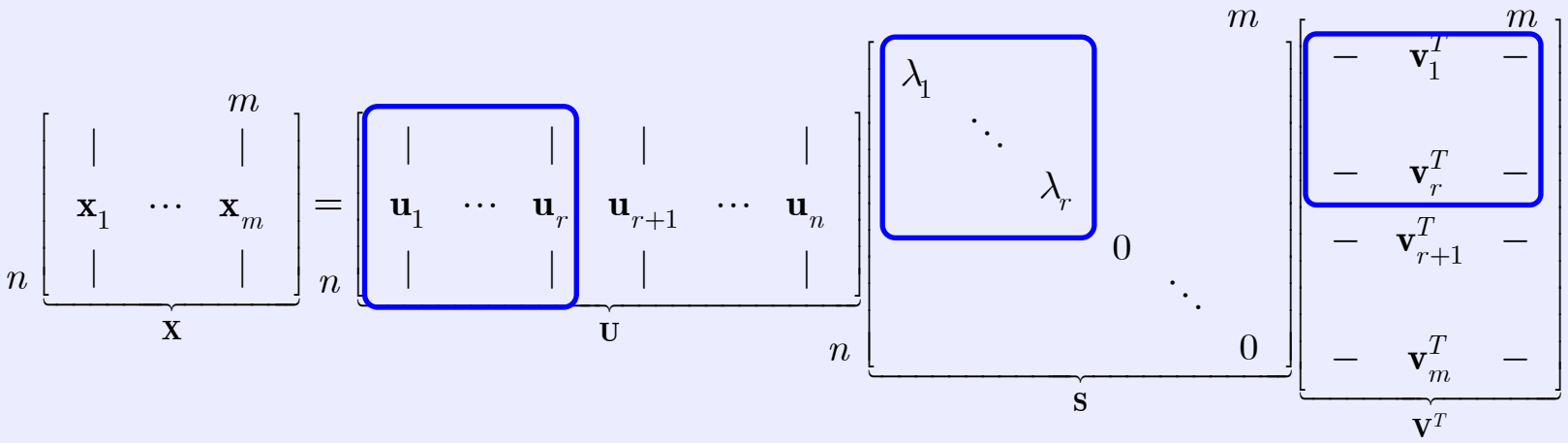
```
>> X*X'
ans =
    11     1
     1    11

>> X'*X
ans =
    10     0     2
     0    10     4
     2     4     2
```

Singular value decomposition (SVD)

- **Economy sized SVD or thin SVD**
  - Avoids computing unnesessary components

$$\underbrace{\mathbf{X}}_{n\times m} = \underbrace{\mathbf{U}}_{n\times n}\ \underbrace{\mathbf{S}}_{n\times m}\ \underbrace{\mathbf{V}}_{m\times m}{}^{T} \longrightarrow \underbrace{\mathbf{X}}_{n\times m} = \underbrace{\tilde{\mathbf{U}}}_{n\times r}\ \underbrace{\tilde{\mathbf{S}}}_{r\times r}\ \underbrace{\tilde{\mathbf{V}}}_{r\times m}{}^{T}$$



Since there is at most $r \le \min\{n, m\}$ non-zero singular values.

- **Truncated SVD**
  - Shrink with $k < r$:

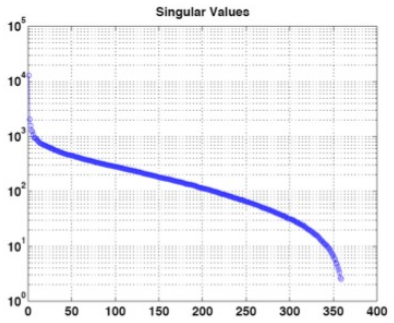
$$\underbrace{\mathbf{X}}_{n\times m} = \underbrace{\mathbf{U}}_{n\times n}\ \underbrace{\mathbf{S}}_{n\times m}\ \underbrace{\mathbf{V}}_{m\times m}{}^{T} \longrightarrow \underbrace{\mathbf{X}}_{n\times m} = \underbrace{\tilde{\mathbf{U}}}_{n\times k}\ \underbrace{\tilde{\mathbf{S}}}_{k\times k}\ \underbrace{\tilde{\mathbf{V}}}_{k\times m}{}^{T}$$

Singular value decomposition (SVD)

- **Spectral view on SVD**
  - Spectrum of $\mathbf{X}$ $\{\lambda_i\}_{i=1}^{r}$ decays very quickly and these singlue values are ordered

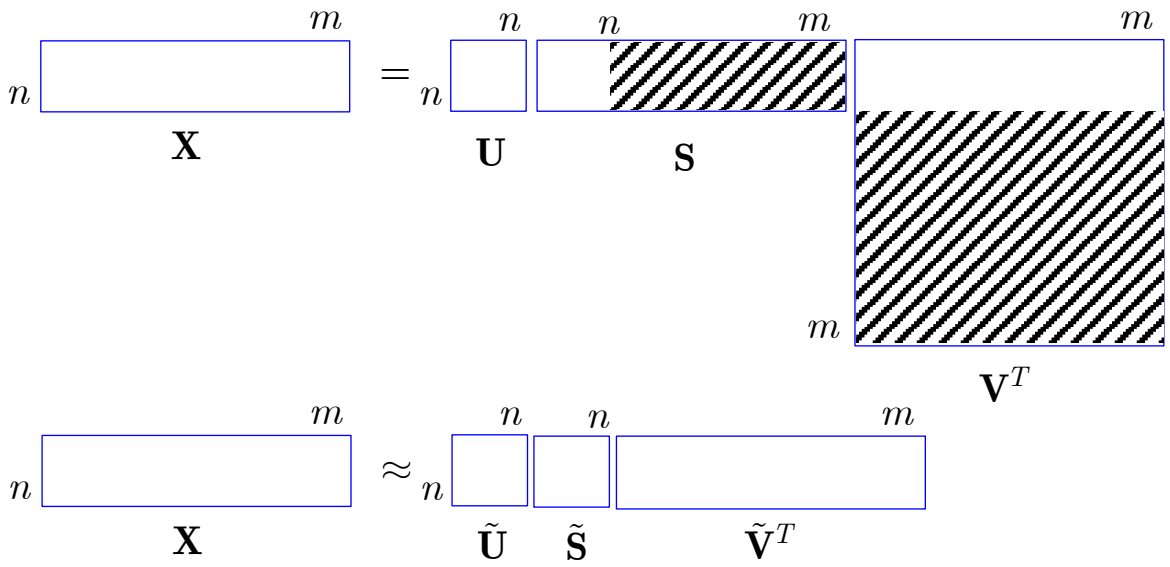
$$\underbrace{\mathbf{X}}_{n\times m} = \underbrace{\tilde{\mathbf{U}}}_{n\times k}\ \underbrace{\tilde{\mathbf{S}}}_{k\times k}\ \underbrace{\tilde{\mathbf{V}}}_{k\times m}^{T} = \sum_{i=1}^{k}\lambda_i \mathbf{u}_i \mathbf{v}_i^T$$

$$\mathbf{X} = \underbrace{\begin{bmatrix}\lambda_1\end{bmatrix}\begin{bmatrix}|\\ \mathbf{u}_1\\ |\end{bmatrix}\begin{bmatrix}- & \mathbf{v}_1^T & -\end{bmatrix}}_{n\times m} + \underbrace{\begin{bmatrix}\lambda_2\end{bmatrix}\begin{bmatrix}|\\ \mathbf{u}_2\\ |\end{bmatrix}\begin{bmatrix}- & \mathbf{v}_2^T & -\end{bmatrix}}_{n\times m} + \cdots + \underbrace{\begin{bmatrix}\lambda_k\end{bmatrix}\begin{bmatrix}|\\ \mathbf{u}_k\\ |\end{bmatrix}\begin{bmatrix}- & \mathbf{v}_k^T & -\end{bmatrix}}_{n\times m}$$



Singular value decomposition (SVD)

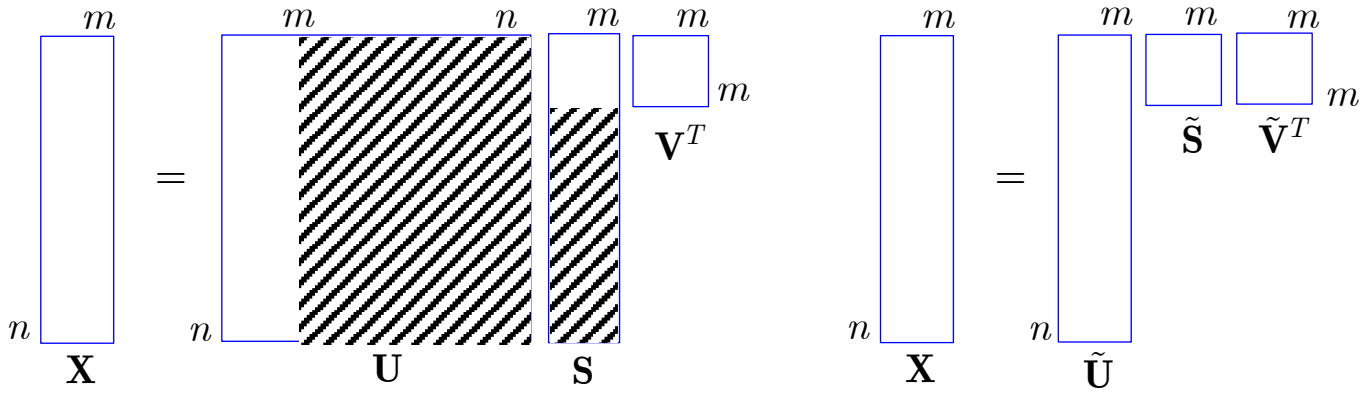
- **Case $m > n$: economy SVD**



$$\mathbf{U} = \tilde{\mathbf{U}} \qquad \tilde{\mathbf{U}}^T\tilde{\mathbf{U}} = \tilde{\mathbf{U}}\tilde{\mathbf{U}}^T = \mathbf{I}_n$$
$$\tilde{\mathbf{V}}^T\tilde{\mathbf{V}} = \mathbf{I} \neq \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T$$

Singular value decomposition (SVD)

- **Case $n > m$: economy SVD**

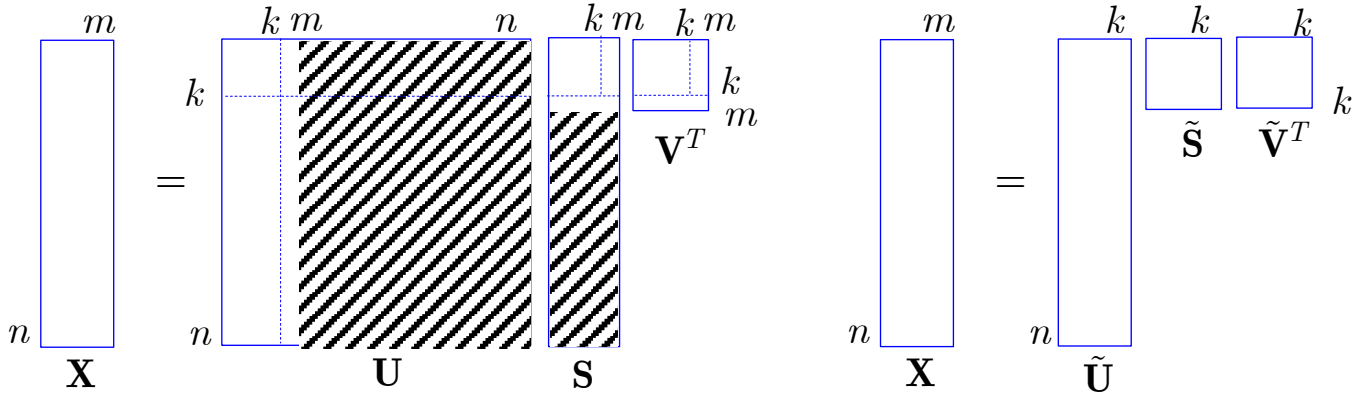


$$\mathbf{V} = \tilde{\mathbf{V}} \qquad \tilde{\mathbf{V}}^T\tilde{\mathbf{V}} = \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T = \mathbf{I}_m$$

$$\tilde{\mathbf{U}}^T\tilde{\mathbf{U}} = \mathbf{I} \neq \tilde{\mathbf{U}}\tilde{\mathbf{U}}^T$$

Singular value decomposition (SVD)

- **Case $k < \min\{m, n\}$: truncated SVD**



$$\tilde{\mathbf{V}}^T\tilde{\mathbf{V}} = \mathbf{I} \neq \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T$$
$$\tilde{\mathbf{U}}^T\tilde{\mathbf{U}} = \mathbf{I} \neq \tilde{\mathbf{U}}\tilde{\mathbf{U}}^T$$

Notations

- **Characteristic function (CF)** of pdf is defined as:

$$\Phi(t) = \int_{-\infty}^{\infty} f(x) e^{jtx} dx = \mathbb{E}_{f(x)} \left[e^{jtX} \right]$$

where $j = \sqrt{-1}$

- The pdf can be recovered as:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi(t) e^{-jtx} dt$$

Moment-generating function of pdf:

$$M(t) = \mathbb{E}_{f(x)}[e^{tX}]$$

Appendix

- Recall D: moments of pdf and cdf

Notations

- Taylor's series expansion: $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$

$$\exp(tX) = \sum_{n=0}^{\infty} \frac{t^n \cdot X^n}{n!} \qquad \exp(jtX) = \sum_{n=0}^{\infty} \frac{(jt)^n \cdot X^n}{n!}$$

$$\Phi(t) = \mathbb{E}_{f(x)}[\exp(jtX))] = 1 + \frac{jt\mathbb{E}_{f(x)}[X]}{1} - \frac{t^2\mathbb{E}_{f(x)}[X^2]}{2!} + \ldots + \frac{(jt)^n\mathbb{E}_{f(x)}[X^n]}{n!}$$

$$M(t) = \mathbb{E}_{f(x)}[\exp(tX))] = 1 + \frac{t\mathbb{E}_{f(x)}[X]}{1} + \frac{t^2\mathbb{E}_{f(x)}[X^2]}{2!} + \ldots + \frac{t^n\mathbb{E}_{f(x)}[X^n]}{n!}$$

The first four moments define the mean, variance, skewness, and kurtosis of the PDF $f(x)$, respectively.

Notations

- Practical computation of moments from i.i.d. sequence $\mathbf{x} = (x_1, \ldots, x_N)$ drawn from $f(x)$
  - The nth **empiral pdf moment estimation**

$$\hat{\mathrm{m}}_n = \frac{1}{N} \sum_{i=1}^{N} x_i^n, \quad n \geq 1$$

  - Which is an unbiased estimation of **true pdf moments**

$$\mathrm{m}_n = \mathbb{E}[X^n] = \int_{-\infty}^{\infty} f(x) x^n dx$$

  - **Moments of CF**:

$$\mathbf{M}_n = \int_{-\infty}^{\infty} \Phi(t) t^n dt$$

Link to pdf: $\mathbf{M}_n = j^n 2\pi \frac{d^n}{dx^n} f(x) \Big|_{x=0}$

Notations

- Practical computation of moments from i.i.d. sequence $\mathbf{x} = (x_1, \ldots, x_N)$ drawn from $f(x)$
  - **Empirical moments of CDF**:
    - Estimate empirical M-bin histogram $\{h(m)\}_{m=0}^{M-1}$ of pdf $f(x)$
    - Estimate K-point CDF with $K = 2^{\lceil \log_2 M \rceil}$ as

$$\Phi(k) = \sum_{m=0}^{M-1} h(m) \exp\left\{\frac{j2\pi mk}{K}\right\}$$

Note that Fourier works in two ways:

$$h(m) = \frac{1}{K} \sum_{k=0}^{K-1} \Phi(k) \exp\left\{-\frac{j2\pi mk}{K}\right\}$$

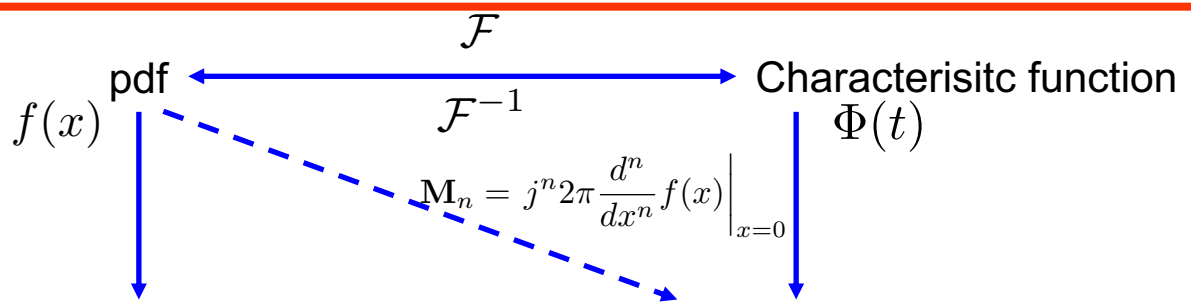
$$0 \le m \le M-1$$

- Compute empirical CDF moments as:

$$\hat{\mathbf{M}}'_n = \sum_{k=0}^{K/2-1} |\Phi(k)| k^n \qquad \hat{\mathbf{M}}_n = \sum_{k=0}^{K-1} \Phi(k) \sin^n\left(\frac{\pi k}{K}\right)$$

Notations

- Summary:



Moments of pdf

$$\mathrm{m}_n = \mathbb{E}[X^n] = \int_{-\infty}^{\infty} f(x) x^n dx$$

Moments of CDF

$$\mathbf{M}_n = \int_{-\infty}^{\infty} \Phi(t) t^n dt$$

Empirical estimation from iid data

$$\hat{\mathrm{m}}_n = \frac{1}{N} \sum_{i=1}^{N} x_i^n, \quad n \geq 1$$

Empirical estimation from iid data

- Estimate empirical M-bin histogram $\{h(m)\}_{m=0}^{M-1}$

$K = 2^{\lceil \log_2 M \rceil}$ → Empirical CDF $\{\Phi(k)\}_{k=0}^{K-1}$

$$\hat{\mathbf{M}}'_n = \sum_{k=0}^{K/2-1} |\Phi(k)| k^n$$

Notations

Applications:

- Outlier detection
- Fake detection
- Secret message detection in images (steganalysis)